Vol. IV 1933

THE ANNALS
of
MATHEMATICAL STATISTICS

Published and Lithoprinted *by*
EDWARDS BROTHERS, INC.
ANN ARBOR, MICH.

THE ANNALS OF MATHEMATICAL STATISTICS

VOL. 4 MAY 1933 NO. 2

CONTENTS

PUBLISHED QUARTERLY BY
AMERICAN STATISTICAL ASSOCIATION
***Publication Office*—Edwards Brothers, Inc., Ann Arbor, Michigan**
***Business Office*—530 Commerce Bldg., New York Univ., New York, N. Y**

Entered as second class matter at the Postoffice at Ann Arbor, Mich., under the Act of March 3rd, 1879.

A quarterly publication of the American Statistical Association, devoted to the theory and application of Mathematical Statistics.

Six dollars per annum.

Reprints of any article in this volume may be obtained at any time from the Editor at the following rates, postage included.

Number of copies	Cost per page
1- 4	2 cents
5-24	1½ cents
25-49	1 cent
50 and over	¾ cent

ADDRESS: Editor, Annals of Mathematical Statistics
Post Office Box 171, Ann Arbor, Michigan

A METHOD OF DETERMINING THE CONSTANTS IN THE BIMODAL FOURTH DEGREE EXPONENTIAL FUNCTION

By

A. L. O'TOOLE

In a paper in this Journal[1] the present writer has discussed some of the mathematical properties of a class of definite integrals which arise in the study of the frequency function

$$(1) \qquad y = e^{-a^2(x^4+p_1x^3+p_2x^2+p_3x+p_4)},\ a \neq 0.$$

This function defines the system of frequency curves for which the method of moments is the best method of fitting[2]—i.e. best in the sense of maximum likelihood—and this fact gives importance to its study. The curves are typically bimodal, the nature and location of the modes being given by the roots of the equation

$$(2) \qquad 4x^3+3p_1x^2+2p_2x+p_3 = 0.$$

The first problem which arose was that of finding an expression for the value of the definite integral

$$(3) \qquad I_0 = \int_{-\infty}^{\infty} e^{-a^2(x^4+p_1x^3+p_2x^2+p_3x+p_4)}dx.$$

If x is replaced by $x - \frac{p_1}{4}$ this integral becomes

$$(4) \qquad I_0 = \int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx+r)}dx,$$

[1] On the system of curves for which the method of moments is the best method of fitting. Vol. IV, No. 1, Feb. 1933, p. 1.

[2] R. A. Fisher, On the mathematical foundations of theoretical statistics, Philosophical Transactions of the Royal Society of London, vol. 222, series A (1921), p. 355.

or

$$(5)\quad I_0 = k\int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx)}dx \quad \text{where} \quad k = e^{-a^2r},$$

or, replacing $x\sqrt{a}$ by x where a is the positive square root of a^2

$$(6)\qquad I_0 = \frac{k}{\sqrt{a}}\int_{-\infty}^{\infty} e^{-(x^4+apx^2+a^{\frac{3}{2}}qx)}dx,$$

or

$$(7)\qquad I_0 = K\int_{-\infty}^{\infty} e^{-(x^4+Px^2+Qx)}dx,$$

where $\quad K = \frac{k}{\sqrt{a}}, \quad P = ap, \quad Q = a^{\frac{3}{2}}q.$

No real loss of generality is incurred in studying (5), (6) or (7) rather than (3). For the purposes of the previous paper it was found convenient to discuss certain special cases of (7) first, then (7) itself and later (5). Having in mind the practical purposes of this note, however, attention will be focused first on the form (5) and afterwards on (3). The transformations from the expressions obtained in the previous paper are very simple. For (5) the special cases studied and a few of the more important results obtained may be stated here as follows:
Type I:

$$p = q = 0.$$

$$I_0 = k\int_{-\infty}^{\infty} e^{-a^2x^4}dx = \frac{k}{\sqrt{2a}}\Gamma\left(\frac{1}{4}\right),$$

$$I_{2n} = k\int_{-\infty}^{\infty} x^{2n}e^{-a^2x^4}dx = \frac{k}{2a^{(2n+1)/2}}\Gamma\left(\frac{2n+1}{4}\right), \quad n = 0, 1, 2, 3, \ldots,$$

$$I_{2n+1} = k\int_{-\infty}^{\infty} x^{2n+1}e^{-a^2x^4}dx = 0, \quad n = 0, 1, 2, 3, \cdots\cdots,$$

$$u_1 = \frac{I_1}{I_0} = 0,$$

$$u_2 = \frac{I_2}{I_0} = \frac{\Gamma(\frac{3}{4})}{a\Gamma(\frac{1}{4})},$$

$$u_3 = \frac{I_3}{I_0} = 0,$$

$$u_4 = \frac{I_4}{I_0} = \frac{\Gamma(\frac{5}{4})}{a^2\Gamma(\frac{1}{4})} = \frac{1}{4a^2},$$

hence

$$a^2 = \frac{1}{4u_4},$$

$$u_{2n} = \frac{I_{2n}}{I_0} = \frac{\Gamma(\frac{2n+1}{4})}{a^n\Gamma(\frac{1}{4})}, \quad n = 0, 1, 2, 3, \ldots\ldots,$$

$$u_{2n+1} = \frac{I_{2n+1}}{I_0} = 0, \quad n = 0, 1, 2, 3, \ldots\ldots .$$

Obviously, of course, k depends upon the total frequency and hence if the total frequency is

$$k = \frac{N}{\int_{-\infty}^{\infty} e^{-a^2x^4}dx} = \frac{2N\sqrt{a}}{\Gamma(\frac{1}{4})}.$$

This curve has a single mode located at $x = 0$ and is symmetrical with respect to the ordinate at $x = 0$.

Type II:

$$q=0,\ p=-2b,\ b>0,$$

$$I_0=k\int_{-\infty}^{\infty}e^{-a^2(x^4-2bx^2)}dx$$

$$=\frac{k}{\sqrt{a}}\int_{-\infty}^{\infty}e^{-(x^4-2abx^2)}dx$$

$$=\frac{k}{\sqrt{a}}\Big[\tfrac{1}{2}\Gamma(\tfrac{1}{4})(1+\frac{a^2b^2}{2!}+\frac{5\cdot1a^4b^4}{4!}+\frac{9\cdot5\cdot1a^6b^6}{6!}+\frac{13\cdot9\cdot5\cdot1a^8b^8}{8!}+\cdots)$$

$$+ab\Gamma(\tfrac{3}{4})(1+\frac{3a^2b^2}{3!}+\frac{7\cdot3\cdot a^4b^4}{5!}+\frac{11\cdot7\cdot3a^6b^6}{7!}+\cdots\cdot)\Big]$$

$$=\frac{k}{\sqrt{a}}\Big[e^{\frac{a^2b^2}{2}}\tfrac{1}{2}\Gamma(\tfrac{1}{4})(1+\frac{a^4b^4}{3\cdot4}+\frac{a^8b^8}{7\cdot3\cdot4^2\cdot2!}+\cdots\cdots)$$

$$+ab\Gamma(\tfrac{3}{4})(1+\frac{a^4b^4}{5\cdot4}+\frac{a^8b^8}{9\cdot5\cdot4^2\cdot2!}+\cdots\cdots)\Big]$$

It was shown that this integral could be expressed in terms of the Bessel functions $J_{\frac{1}{4}}$ and $J_{-\frac{1}{4}}$ as follows:

$$I_0=\frac{k}{\sqrt{a}}\left[\left(\frac{a^2b^2}{2}\right)^{\frac{1}{4}}e^{\frac{a^2b^2}{2}}\left\{AJ_{\frac{1}{4}}\left(\frac{-ia^2b^2}{2}\right)+BJ_{-\frac{1}{4}}\left(\frac{-ia^2b^2}{2}\right)\right\}\right]$$

where

$$A=\frac{2^{\frac{3}{4}}\Gamma(\frac{1}{4})\Gamma(\frac{3}{4})}{\sqrt[4]{-i}},$$

$$B=\frac{\frac{1}{2}\Gamma(\frac{1}{4})\Gamma(\frac{3}{4})}{\sqrt[4]{2i}},\quad i=\sqrt{-1}.$$

If the total frequency is N

$$k=\frac{N}{\int_{-\infty}^{\infty}e^{-a^2(x^4-2bx^2)}dx}$$

This curve is symmetrical with respect to the ordinate at $x=0$ and has two real modes located at $x=\pm\sqrt{b}$.

Type III: $p=0$.

$$I_o=k\int_{-\infty}^{\infty}e^{-a^2(x^4+qx)}dx$$

$$=\frac{k}{\sqrt{a}}\int_{-\infty}^{\alpha}e^{-(x^4+a^{\frac{3}{2}}qx)}dx$$

$$=\frac{k}{2\sqrt{a}}\sum_{n=0}^{\infty}\frac{(a^{\frac{3}{2}}q)^{2n}}{(2n)!}\Gamma\left(\frac{2n+1}{4}\right)$$

$$=\frac{k}{2\sqrt{a}}\left[\Gamma\left(\frac{1}{4}\right)\left\{1+\frac{(a^{\frac{3}{2}}q)^4}{4\cdot 4!}+\frac{5(a^{\frac{3}{2}}q)^8}{4^2\cdot 8!}+\frac{9\cdot 5(a^{\frac{3}{2}}q)^{12}}{4^3\cdot 12!}+\cdots\right\}\right.$$
$$\left.+\Gamma\left(\frac{3}{4}\right)\left\{\frac{(a^{\frac{3}{2}}q)^2}{2!}+\frac{3(a^{\frac{3}{2}}q)^6}{4.6!}+\frac{7\cdot 3(a^{\frac{3}{2}}q)^{10}}{4^2\cdot 10!}+\cdots\right\}\right],$$

$$k=\frac{N}{\int_{-\infty}^{\infty}e^{-a^2(x^4+qx)}dx}$$

This curve is not symmetrical and has only one real mode, that mode being located at x equal to the real cube root of negative q.

Type IV: The general case.

$$I_0 = k\int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx)}dx$$

$$= \frac{k}{a}\int_{-\infty}^{\infty} e^{-(x^4+apx^2+a^{\frac{3}{2}}qx)}dx$$

$$= K\int_{-\infty}^{\infty} e^{-(x^4+Px^2+Qx)}dx.$$

It was shown that the value of this integral could be expressed as an infinite series each term of which involved two Bessel functions. But, as pointed out near the close of the previous paper, although this infinite series may be considered a theoretical solution of the problem, it does not lead to a simple method of determining the constants a^2, p, q, k which appear in the frequency function. It is the purpose of this note to give a practical method of determining these constants.

Beginning with (5)

$$I_0 = k\int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx)}dx,$$

the *n-th* moment u_n' is defined by

$$(8) \qquad u_n' = \frac{k\int_{-\infty}^{\infty} x^n e^{-a^2(x^4+px^2+qx)}dx}{k\int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx)}dx}$$

$$= \frac{k}{I_0}\int_{-\infty}^{\infty} x^n e^{-a^2(x^4+px^2+qx)}dx.$$

Integrate I_0 by parts, letting $u = e^{-a^2(x^4+px^2)}$ and $dv = e^{-a^2qx}dx$. Then

$$(9) \qquad I_0 = -\frac{k}{q}\int_{-\infty}^{\infty}(4x^3+2px)e^{-a^2(x^4+px^2+qx)}dx.$$

Divide by I_0 and multiply by q and the result is

$$(10) \qquad q = -(4u_3' + 2pu_1').$$

Start again with I_0 in the form (5) and integrate by parts letting

$u = e^{-a^2(x^4+px^2+qx)}$ and $dv = dx$. Then

$$(11) \qquad I_0 = ka^2\int_{-\infty}^{\infty}(4x^4+2px^2+qx)e^{-a^2(x^4+px^2+qx)}dx.$$

Divide by I_0 and then

$$1 = a^2(4u_4' + 2pu_2' + qu_1')$$

or

$$(12) \qquad a^2 = \frac{1}{4u_4' + 2pu_2' + qu_1'}\,.$$

Now integrate (11) by parts with $u = e^{-a^2(x^4+px^2+qx)}$ and $dv = (4x^4+2px^2+qx)\,dx$. This leads to

$$(13) \qquad \begin{aligned} I_0 = \frac{ka^4}{30}\int_{-\infty}^{\infty}(96x^8&+128px^6+84qx^5+40p^2x^4 \\ &+50pqx^3+15q^2x^2)e^{-a^2(x^4+px^2+qx)}dx. \end{aligned}$$

Divide by I_0 and obtain

$$1 = \frac{a^4}{30}(96u_8' + 128pu_6' + 84qu_5' + 40p^2u_4' + 50pqu_3' + 15q^2u_2')$$

or

$$(14)\quad a^4 = \frac{30}{96u'_8 + 128pu'_6 + 84qu'_5 + 40p^2u'_4 + 50pqu'_3 + 15q^2u'_2}.$$

Squaring in (12) the result is

$$(15)\quad a^4 = \frac{1}{(4u'_4 + 2pu'_2 + qu'_1)^2}$$

$$= \frac{1}{16u_4'^2 + 4p^2u_2'^2 + q^2u_1'^2 + 16pu'_2u'_4 + 8qu'_1u'_4 + 4pqu'_1u'_2}.$$

Eliminating a^4 between (14) and (15) the equation

$$(16)\quad p^2(40u'_4 - 120u_2'^2) + q^2(15u'_2 - 30u_1'^2) + pq(50u'_3 - 120u'_1u'_2)$$
$$+ p(128u'_6 - 480u'_2u'_4) + q(84u'_5 - 240u'_1u'_4) + (96u'_8 - 48u_4'^2) = 0.$$

Using relation (10) and

$$(17)\quad q^2 = 16u_3'^2 + 16pu'_1u'_3 + 4p^2u_1'^2$$

eliminate q from (16) obtaining

$$(18)\quad 5Ap^2 + 2Bp + 2C = 0$$

and hence

$$(19)\quad p = \frac{-B \pm \sqrt{B^2 - 10AC}}{5A}$$

where

$$(20)\quad \begin{cases} A = 2u'_4 - 6u'_2 + 15u_1'^2u'_2 - 6u_1'^4 - 5u'_1u'_3, \\ B = 90u'_1u'_2u'_3 - 60u_1'^3u'_3 - 25u_3'^2 + 16u'_6 - 60u'_2u'_4 - 21u'_1u'_5 + 60u_1'^2u'_4, \\ C = 30u'_2u_3'^2 - 60u_1'^2u_3'^2 - 42u'_3u'_5 + 120u'_1u'_3u'_4 + 12u'_8 - 60u_4'^2. \end{cases}$$

In order to decide upon one of the two values of p furnished by (18) notice that, equating the first derivative of the frequency function to zero, the location of the two modes and the minimum point between them is determined by the roots of the equation

$$4x^3+2px+q=0. \tag{21}$$

The condition for three real distinct roots in this equation is

(22) $-8p^3>27q^2$. which requires $p<0$,

where q is found from (17). If $-8p^3=27q^2$ then one of the modes coincides with the minimum point. If $p=q=0$ then both modes coincide with the minimum point.

Extracting the square root in (17) gives two values of q differing only in sign. Now it is easy to show either by geometrical considerations or by examining the algebraic manipulations leading to (18) that p is independent of the sign of q. Changing q to $-q$ in (5) has the same effect as changing x to $-x$ or, that is, reversing the order of the distribution and curve. Also, changing x to $-x$ leaves the even moments unaltered but changes the sign of every odd moment. Hence if the value of the function at the modal position on the left is greater than the value of the function at the modal position on the right then q is greater than zero. And if the value of the function at the modal position on the left is less than the value of the function at the modal position on the right then q is less than zero. If $q=0$ the curve is symmetrical with respect to the ordinate at $x=0$. Hence p and q are determined by (19), (17) and (22), the sign of q being fixed by examination of the data of the problem or, if necessary, by trial. The value of a^2 is then found by taking the positive square root in (15). Of course (14) would give the same value for a^2.

Now that a^2, p and q are determined, there remains only k to be found. If the total frequency is N then

$$k\int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx)}dx=N$$

and hence

$$k = \frac{N}{\int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx)}dx} \tag{23}$$

where the numerical value of the integral in the denominator can be found by mechanical quadrature to any desired degree of approximation. For purposes of the quadrature involved here it will be found that the simple rectangle quadrature formula will give as good results as could be desired.[3] Having found k then the constant r is also known since

$$\left[\begin{array}{l} k = e^{-a^2 r}, \\ \\ r = \dfrac{\log_e k}{-a^2} = \dfrac{\log_{10} k}{-a^2 \log_{10} e} \end{array}\right. . \tag{24}$$

The points of inflexion are located by equating the second derivative of the function to zero. The equation is

$$a^2(4x^3+2px+q)^2-2(6x^2+p)=0. \tag{25}$$

If now x be replaced by $x+m$ then

$$I_0 = \int_{-\infty}^{\infty} e^{-a^2(x^4+px^2+qx+r)}dx \tag{5}$$

becomes

$$I_0 = \int_{-\infty}^{\infty} e^{-a^2(x^4+p_1x^3+p_2x^2+p_3x+p_4)}dx \tag{3}$$

[3] On the degree of Approximation of Certain Quadrature Formulas. Annals of Mathematical Statistics, vol. IV, No. 2, May 1933, p. 143 by A. L. O'Toole.

where

$$(26)\quad \begin{cases} p_1 = 4m \\ p_2 = 6m^2 + p, \\ p_3 = 4m^3 + 2mp + q, \\ p_4 = m^4 + m^2 p + mq + r. \end{cases}$$

The data[4] in the first two columns of the table given here will provide the basis for an illustration of the method described above for determining the constants. The numbers in the first column are the classes into which the plants were divided. In the second column are found the frequencies corresponding to the various classes. In constructing the third column the origin for x was arbitrarily placed to correspond to the class 25. Taking

$$u'_n = \frac{\Sigma x^n f(x)}{\Sigma f(x)}$$

the first six moments and the eighth moment are found to be

$$u'_1 = \frac{88}{452} = 0.1946903,$$

$$u'_2 = \frac{14086}{452} = 31.16372,$$

$$u'_3 = \frac{12248}{452} = 27.09735,$$

$$u'_4 = \frac{1000264}{452} = 2212.973,$$

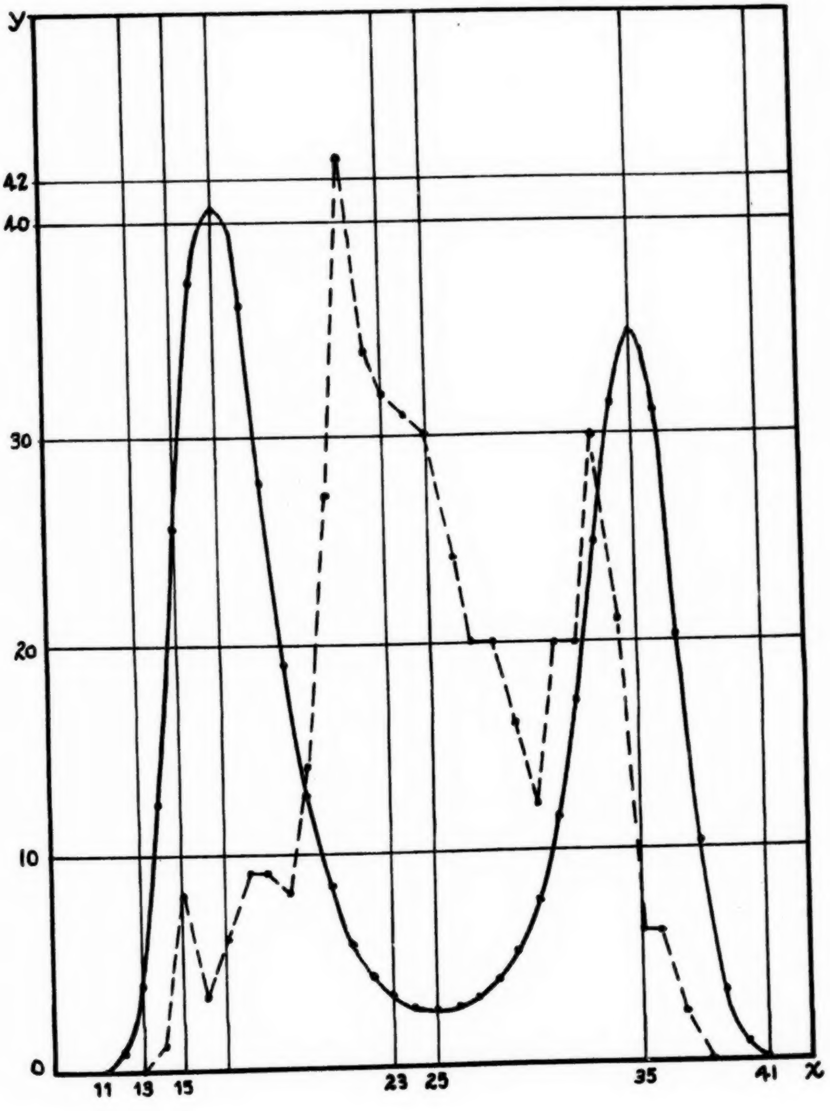
$$u'_5 = \frac{185480}{452} = 410.3540,$$

$$u'_6 = \frac{94571296}{452} = 209228.5,$$

$$u'_8 = \frac{10428472504}{452} = 23071842.$$

[4] This data, except for slight modifications, was extracted from that of W. L. Tower on the Seriation of Counts of Rays of Chrysanthemum Leucanthemum, Biometrika No. 1, 1901-2, p. 313.

CLASS	FREQUENCY $f(x)$	x	y'	$2.5y'=y$	COMPUTED FREQUENCY
9		-16	.031102	.077755	0
10		-15	.300472	.751180	1
11		-14	1.583594	3.958985	4
12	1	-13	4.991068	12.477670	12
13	8	-12	10.246676	25.616690	26
14	3	-11	14.831798	37.079495	37
15	6	-10	16.280112	40.700280	41
16	9	- 9	14.482540	36.206350	36
17	9	- 8	11.089036	27.722590	28
18	8	- 7	7.712195	19.280487	19
19	14	- 6	5.108903	12.772257	13
20	27	- 5	3.359072	8.397680	8
21	43	- 4	2.269766	5.674415	6
22	34	- 3	1.621764	4.054410	4
23	32	- 2	1.252760	3.131900	3
24	31	- 1	1.062910	2.657275	3
25	30	0	1.000000	2.500000	3
26	24	1	1.046530	2.616325	3
27	20	2	1.214445	3.036112	3
28	20	3	1.547935	3.869837	4
29	16	4	2.133051	5.332627	5
30	12	5	3.108096	7.770240	8
31	20	6	4.654336	11.635840	12
32	20	7	6.917724	17.294310	17
33	30	8	9.793409	24.483522	24
34	21	9	12.593310	31.483275	31
35	6	10	13.938151	34.845377	35
36	6	11	12.502565	31.256412	31
37	2	12	8.504388	21.260970	21
38		13	4.078577	10.196442	10
39		14	1.274131	3.185327	3
40		15	.238029	.595072	1
41		16	.024259	.060647	0
	452		180.792704		452


y
42
40
30
20
10
0
11
13
15
23
25
35
41
x

Formulas (20) give

$$A = -1409.786$$

$$B = -790428.9,$$

$$C = -15354106.$$

Hence from (19)

$$p = -202.7862$$

or

$$p = -21.48292.$$

But $p = -21.48292$ and the value of q to which it leads do not satisfy the relation (22) hence use $p = -202.7862$. Calculate q from (17) and use the positive square root since an examination of the data shows that the value of the function at the left modal value is greater than the value of the function at the right modal value. Hence

$$q = 29.4284.$$

Formula (15) now gives as the positive square root

$$a^2 = 0.0002640.$$

Using these values for a^2, p, q the values of the function

$$y' = e^{-a^2(x^4+px^2+qx)}$$

are calculated for integral values of x from $x = -16$ to $x = 16$ and tabulated in column four. The constant k is then found by dividing the total frequency 452 by the sum of column four. Hence

$$k = \frac{452}{180.792704} = 2.500100.$$

By (24)

$$r = -3472.578.$$

The function can now be written

$y = 2.5e^{-.0002640(x^4-202.7862x^2+29.4284x)}$ taking $k = 2.5$,

or $y = e^{-.0002640(x^4-202.7862x^2+29.4284x-3472.578)}$.

The values of the ordinates for this function are given in column five and to the nearest integer in column six.

Equation (21) becomes

$$4x^3-2(202.7862)x+29.4284=0$$

which has the roots (approximate) $x=-10.1$, $x=0.07$, $x=10.03$ It should be noted that the sum of the three roots must equal zero. Hence the modes are located at $x=-10.1$ and at $x=10.03$ with the minimum point at $x=0.03$. These roots can be determined to any desired number of decimal places by Horner's method.

If now x is replaced by $x=-25$ so that the new values of x are respectively equal to the numbers in the class column, the function becomes

$$y=e^{-.0002640(x^4-100x^3+3547.214x^2-52331.26x+259605.3)}$$

The modes are now located at $x=14.9$ and $x=35.03$ with the minimum point at $x=25.07$. In the figure are shown the original distribution and the curve represented by this equation.

A. L. O'Toole

ON THE TCHEBYCHEF INEQUALITY OF BERNSTEIN

By

CECIL C. CRAIG[1]

From Tchebychef's inequality we know that if $x_1, x_2, \cdots, x_n$ are a set of independent statistical variables with

$$m_{x_1} = m_{x_2} = \cdot\cdot\cdot\cdot\cdot = m_{x_n} = 0,$$

and

$$\sigma^2 = \sigma^2_{x_1} + \sigma^2_{x_2} + \cdot\cdot\cdot\cdot\cdot + \sigma^2_{x_n},$$

then the probability P that

$$-t\sigma \leq x_1 + x_2 + \cdots + x_n \leq t\sigma$$

satisfies the inequality,

$$P \geq 1 - \frac{1}{t^2} \cdot$$

This gives a lower limit for P which is often unsatisfactory. Improvement of this result requires further hypotheses. As is well-known, Pearson, Camp, Guldberg, Meidel, Narumi,[2] and Smith[3] have attacked this problem with considerable success. Another interesting and important attempt in this direction due to S. Bernstein seems to have generally escaped attention in the English-speaking world, at least, since it has been published only in Russian.[4] Because of the latter fact, it seems necessary to give

[1] This paper was written in substantially its present form during the author's tenure of a National Research Fellowship at Stanford University.

[2] For references to all these papers except Smith's and a brief discussion see Rietz, H. L., Mathematical Statistics, (Open Court Publishing Company, Chicago, 1927), pp. 140-144.

[3] Smith, C. D., On Generalized Tchebychef Inequalities in Mathematical Statistics, American Journal of Mathematics, Vol. 52, (1930), pp. 109-126.

[4] Bernstein, S., Theory of Probability, (Moscow, 1927), pp. 159-165. The present account of this work of Bernstein is taken from a lecture of Professor J. V. Uspensky.

a brief account of this work of Bernstein's preliminary to the remarks based on it the writer wishes to make.

Bernstein imposed the condition in addition that

(1) $E(|x_i|^k) \leq \frac{\sigma_{x_i}^2}{2} k!\, h^{k-2}; \; k \geq 2, \quad i = 1, 2, \cdots\cdots, n,$

$(E(x)$ is read "the mathematical expectation of x.") in which h is arbitrary. (This condition is satisfied, e.g., if the x_i's are bounded.) and used the following lemma due to Tchebychef. Let the statistical variable u be always > 0. If $E(u) = A$, then the probability Q that $u \geq At^2$ satisfies the inequality, $Q \leq \frac{1}{t^2}$

Then taking,

$$u = e^{\varepsilon(x_1 + x_2 + \cdots\cdots + x_n)},$$
$$= e^{\varepsilon x_1} e^{\varepsilon x_2} \cdots\cdots e^{\varepsilon x_n},$$

in which ε is arbitrary,

$$E(u) = E(e^{\varepsilon x_1}) E(e^{\varepsilon x_2}) \cdots\cdots E(e^{\varepsilon x_n}).$$

Now

$$e^{\varepsilon x_i} = 1 + \varepsilon x_i + \frac{\varepsilon^2 x_i^2}{2!} + \frac{\varepsilon^3 x_i^3}{3!} + \cdots\cdots,$$

and under the condition (1),

$$E(e^{\varepsilon x_i}) \leq 1 + \frac{\varepsilon^2 \sigma_{x_i}^2}{2} + \frac{\varepsilon^3 \sigma_{x_i}^2 h}{2} + \frac{\varepsilon^4 \sigma_{x_i}^2 h^2}{2} + \cdots\cdots$$

If it is assumed that

$$|\varepsilon| h \leq c < 1$$

then

$$E(e^{\varepsilon x_i}) \leq 1 + \frac{\varepsilon^2 \sigma_{x_i}^2}{2(1-c)} < e^{\frac{\varepsilon^2 \sigma_{x_i}^2}{2(1-c)}},$$

and thus

(2) $$E(u) < e^{\frac{\varepsilon^2 \sigma^2}{2(1-c)}}$$

If in the inequality, $u \geq At^2$, a greater quantity is substituted for A , then certainly $Q \leq \frac{1}{t^2}$. Therefore the probability Q of

$$u \geq e^{\frac{\varepsilon^2\sigma^2}{2(1-c)}} e^{\tau^2}$$

satisfies the inequality

$$Q \leq e^{-\tau^2}$$

Now

$$u = e^{\varepsilon(x_1 + x_2 + \cdots\cdot x_n)} \geq e^{\tau^2 + \frac{\varepsilon^2\sigma^2}{2(1-c)}}$$

implies for $\varepsilon > 0$,

$$x_1 + x_2 + \cdots + x_n \geq \frac{\tau^2}{\varepsilon} + \frac{\varepsilon\sigma^2}{2(1-c)} .$$

The value of ε is next chosen so as to make Q a minimum, i.e., so as to make $\frac{\tau^2}{\varepsilon} + \frac{\varepsilon\sigma^2}{2(1-c)}$ a minimum. Thus

$$\varepsilon^2 = \frac{2(1-c)\tau^2}{\sigma^2} .$$

Then the probability Q that

$$x_1 + x_2 + \cdots + x_n \geq \tau\sigma\left(\frac{2}{1-c}\right)^{\frac{1}{2}}$$

satisfies the inequality,

$$Q \leq e^{-\tau^2}$$

if $\varepsilon^2 = \frac{2(1-c)\tau^2}{\sigma^2}$; $\varepsilon \leq \frac{c}{n}$ with $c < 1.$

To get the corresponding result for the lower limit of the sum $x_1 + x_2 + \cdots\cdots + x_n$, it is only necessary to choose $\varepsilon < 0$ and as before, the probability, Q' , that

$$x_1 + x_2 + \cdots + x_n \leq -\tau\sigma\left(\frac{2}{1-c}\right)^{\frac{1}{2}}$$

satisfies the inequality,

$$Q' \leq e^{-\tau^2}$$

if also $\varepsilon^2 = \frac{2(1-c)}{\sigma^2}\tau^2$ and $|\varepsilon| \leq \frac{c}{h}$ with $c < 1$.

Combining these two results, if P is the probability of

$$-\tau\sigma\left(\frac{2}{1-c}\right)^{\frac{1}{2}} \leq x_1 + x_2 + \cdots + x_n \leq \tau\sigma\left(\frac{2}{1-c}\right)^{\frac{1}{2}},$$

then since

$$P + Q + Q' = 1,$$

$$P \geq 1 - 2e^{-\tau^2}.$$

But setting

$$\tau\sigma\left(\frac{2}{1-c}\right)^{\frac{1}{2}} = \omega,$$

and also,

$$\varepsilon^2 \leq \frac{c^2}{h^2},$$

the condition

$$\frac{2(1-c)}{\sigma^2}\tau^2 \leq \frac{c^2}{h^2}$$

(Bernstein set $\varepsilon^2 = \frac{c^2}{h^2}$, using merely the equality sign in this condition. The value of c as here given is necessary in the author's developments below.) must be satisfied, or what is the same thing,

$$\frac{2(1-c)^2\omega^2}{2\sigma^4} \leq \frac{c^2}{h^2},$$

from which

$$c \geq \frac{h\omega}{\sigma^2 + h\omega}.$$

This last quantity on the right is positive and < 1 as required so that the constants can actually be chosen as specified.

This gives

$$\tau = \omega \left[2(\sigma^2 + h\omega)\right]^{-\frac{1}{2}},$$

and finally the probability, P, that

$$-\omega \leq x_1 + x_2 + \cdots\cdots x_n \leq \omega$$

is such that

$$P \geq 1 - 2e^{-\frac{\omega^2}{2\sigma^2 + 2h\omega}},$$

or setting $\omega = t\sigma$

$$P \geq 1 - 2e^{-\frac{t^2}{2 + \frac{2ht}{\sigma}}}. \tag{3}$$

It is to be observed that generally the quantity $\frac{2ht}{\sigma}$ rapidly decreases with increasing n.

This is the inequality reached by Bernstein under the condition (1).

If all the x_i's are bounded, if, say, always

$$|x_i| \leq b, \quad i = 1, 2, \cdots\cdots, n,$$

one may take $h = \frac{b}{3}$.

It is the purpose of the author's remarks to discuss less severe conditions than (1) under which the inequality (3) can be obtained. These more general conditions are obtained, however, at the expense of assuming quite generally satisfied regularity conditions with regard to the "tails" of the frequency distribution of x, which needs not necessarily to be regarded as the sum of n component variables, $x_1, x_2, \cdots\cdots, x_n$.

If we now take

$$u = e^{\varepsilon x} \tag{4}$$

we have

$$E(u) = \int_{-\infty}^{\infty} dF(x) e^{\varepsilon x} \quad (F(x) \text{ is the probability function of } x).$$

$$= \int_{-\infty}^{\infty} dF(x)\left(1 + \varepsilon x + \frac{\varepsilon^2 x^2}{2!} + \frac{\varepsilon^3 x^3}{3!} + \cdots\right).$$

The condition (1) insures that the series under the sign of integration may be integrated over the interval $(-\infty, \infty)$. But the series can also be integrated over the same interval if it converges uniformly in any fixed finite interval, which it does, and if the series $\sum_{n=0}^{\infty} g_n(y)$, where

$$g_n(y) = \int_{-y}^{y} dF(x) \frac{\varepsilon^n x^n}{n!},$$

converges uniformly in the interval $(-\infty, \infty)$.

Formally, at least,

$$(5) \qquad E(u) = 1 + \mu_2 \frac{\varepsilon^2}{2!} + \mu_3 \frac{\varepsilon^3}{3!} + \cdots ,$$

in which μ_k is the k-*th* moment about the mean of x. If

$$(6) \qquad |\mu_k| \leq \frac{k!}{2} \mu_2 h^{k-2}, \; k \geq 2,$$

for some $h > 0$, then for $h\,|\varepsilon| \leq c < 1$ the right hand side of (5) is convergent and is $\leq 1 + \frac{\varepsilon^2 \sigma^2}{2(1-c)}$ as before. Now let us suppose that the condition (6) is satisfied not only for moments taken over the whole interval $(-\infty, \infty)$ but also for moments taken over any interval which includes the interval $(-b, b)$ in which b is an arbitrarily large though finite number. This is the *first regularity condition*, mentioned above, which we shall impose on the tails of the frequency function of x.

Then it is obvious, from the remark above, that

$$\sum_{n=0}^{\infty} q_{n(y)}$$

is uniformly convergent in the interval for $|y| \leq b$ for $\frac{b}{3}|\mathcal{E}| \leq c < 1$ And for $|y| > b$ it is also obvious that for $h|\mathcal{E}| \leq c < 1$,

$$\sum_{n=0}^{\infty} q_n(y)$$

is uniformly convergent if our first regularity condition is satisfied. And since $|\mathcal{E}|$ may be taken arbitrarily small, the inequality (3) follows as before.

It is evident that if our first regularity condition holds, that the condition (6) is more general than the condition (1). And it is easily seen that this first regularity condition holds for a very wide class of frequency functions. For, in order for it to hold, it is sufficient that the frequency curve (continuous or not) outside some finite interval $(-b, b)$ about the mean as center, be never increasing as $|x|$ increases and that if $f(x)$ be the ordinate of the frequency curve at the abscissa x. always $f(x) \geq f(-x)$ or else always $f(x) \leq f(-x)$ for $x > b$.

But if the first regularity condition be satisfied, then for all intervals which include $(-b, b)$ the corresponding moments have upper limits in absolute value. And if this be so for all such intervals, the semi-invariants (of Thiele) will also have upper limits for their absolute values. If λ_k is the k-th semi-invariant, we will take for our *second regularity condition* on the tails of the frequency distribution of x, that

$$(7) \qquad |\lambda_k| \leq \frac{k!}{2} \lambda_2 h^{k-2} \qquad k \geq 2 \quad (\lambda_2 = \mu_2)$$

for some $h > 0$ if λ_k is taken for any interval which includes the arbitrarily large, though finite, interval $(-b, b)$.

If this second regularity condition holds, it is again easy to show that (5) is an equality if $h|\mathcal{E}| \leq c < 1$. The right member of (5) is still uniformly convergent in the interval $(-b,b)$ for $\frac{b}{3}\,|\varepsilon| \le c < 1$. For all intervals which include $(-b,b)$ we use the formal identity which defines the semi-invariants of Thiele:

$$(8) \quad e^{\lambda_2\frac{\varepsilon^2}{2!} + \frac{1}{3!}\lambda_3\frac{\varepsilon^3}{3!} + \cdots} = 1 + \mu_2\frac{\varepsilon^2}{2!} + \mu_3\frac{\varepsilon^3}{3!} + \cdots = e^{\phi(\varepsilon)}$$

Under the condition (7), $\phi(\varepsilon)$ is uniformly convergent over the intervals in question for $h|\varepsilon| \le c < 1$ and for these values of ε, (8) becomes an equality since its second member is only the first arranged in powers of ε. Moreover, on account of (7) the right member must be uniformly convergent for all intervals which include $(-b,b)$.

At least one important class of frequency distributions satisfies our second regularity condition. The distributions of characteristics in samples of N have finite ranges as long as N is finite and they commonly have semi-invariants which are rapidly decreasing with increasing N. If such distributions approach normality their semi-invariants of order above the second approach zero, in particular they may become in absolute value less than or equal to the corresponding semi-invariants of a Pearson's Type III distribution which are given by

$$\frac{\lambda_k}{\lambda_2^{\frac{k}{2}}} = \frac{(k-1)!}{a^{k-2}} \qquad k \ge 2$$

in which $a = \dfrac{2\lambda_2^{\frac{3}{2}}}{\lambda_3}$, or

$$\lambda_k = (k-1)!\,\lambda_2\left(\frac{\sigma}{a}\right)^{k-2}.$$

Taking $h = \left|\frac{\sigma}{a}\right|$ it is easy to see that such distributions satisfy our second regularity condition. The smaller the skewness of the Type III distribution, the smaller h may be taken. Thus in such cases we can give a lower limit for $P(|x| \le t\sigma)$, the probability that $|x| \le t\sigma$, which is improved with decreasing skewness of the Type III distribution. By the use of the first regularity condition we could only take $h = \frac{\sigma}{2}$ as the distribution approaches normality.

As a second application, let us suppose that $x = x_1 + x_2$ in which x_1 and x_2 are independent, and in which the semi-invariants of the distribution of x_1 are $\ell_2 (= \sigma_1^2), \ell_3, \ell_4 \cdots$, and the semi-invariants of the distribution of x_2 are $\gamma_2 (= \sigma_2^2) \gamma_3, \gamma_4, \cdots$. Then the distribution of x has for semi-invariants

$$\lambda_2 = \ell_2 + \gamma (= \sigma^2), \quad \lambda_3 = \ell_3 + \gamma_3, \quad \lambda_4 = \ell_4 + \gamma_4, \cdots\cdots.$$

Further let it be assumed that $\frac{\sigma_2}{\sigma_1} < 1$, and that the distribution of x_2 satisfies our second regularity condition.

Then

$$P(|x| \le t\sigma) > P(|x_1| \le t\sigma_1)\, P(|x_2| \le t(\sigma - \sigma_1))$$

But

$$P(|x_2| \le t(\sigma-\sigma_1)) = P\left(|x_2| \le \frac{t(\sigma-\sigma_1)}{\sigma_2}\sigma_2\right)$$

$$> 1-2e^{\frac{\frac{-t^2(\sigma-\sigma_1)^2}{\sigma_2^2}}{2+2\frac{h}{\sigma_2}\frac{t(\sigma-\sigma_1)}{\sigma_2}}}$$

Now

$$\frac{\sigma-\sigma_1}{\sigma_2} = \frac{(\sigma_1^2+\sigma_2^2)^{\frac{1}{2}}-\sigma_1}{\sigma_2} < 1 \qquad \left\{ \begin{array}{l} (1+x^2)^{\frac{1}{2}} < 1+x \\ \text{if } 0 < x < 1 \end{array} \right.$$

so that we get

$$P(|x_2| \le t(\sigma-\sigma_1)) > 1-2e^{-\frac{t^2}{2+2h\frac{t}{\sigma_2}}}.$$

This gives finally in such cases

$$P(|x| \le t\sigma) > P(|x_1| \le t\sigma) - 2e^{-\frac{t^2}{2+2h\frac{t}{\sigma_2}}}$$

Cecil C. Craig

ON CORRELATION SURFACES OF SUMS WITH A CERTAIN NUMBER OF RANDOM ELEMENTS IN COMMON*

By

Carl H. Fischer

Introduction. The study of correlation due to a common factor has been a more or less familiar one in the literature of mathematical statistics. Kapteyn,[1] in an exposition of the Pearsonian coefficient of correlation, considered the correlation between two sums of normally distributed variables, the sums having k random elements in common. In 1920, Rietz[2] devised urn schemata which yield sums with common items involved in such a way that the correlation and regression properties can be dealt by a priori methods. In a later paper, Rietz[3] considered two variables, each the sum of two random drawings of elements from a continuous rectangular distribution, with one of the elements in common. Here, the emphasis was placed principally upon the description of the correlation surface. Some other aspects and extensions of this problem were brought out by Karl Pearson[4] in an editorial discussion of Rietz's paper.

In the literature, the theory of correlation has been discussed principally in connection with its applications. One of the objects of some of the above-mentioned papers is the establishment of a closer connection between correlation theory and abstract probability theory. Such a connection would give a more precise

*Presented to the American Mathematical Society, Dec. 28, 1931.

[1] J. C. Kapteyn, "Definition of the Correlation-Coefficient," Monthly Notices of the Royal Astronomical Society, Vol. 72 (1912), pp. 518-525.

[2] H. L. Rietz, "Urn Schemata as a Basis for the Development of Correlation Theory," Annals of Mathematics, Vol. 21 (1920), pp. 306-322.

[3] H. L. Rietz, "A Simple Non-Normal Correlation Surface," Biometrika, Vol. 24 (1932), pp. 288-291.

[4] Karl Pearson, "Professor Rietz's Problem," (Editorial), Biometrika, Vol. 24 (1932), pp. 290-291.

meaning to correlation and would tend to make the study of correlation theory more attractive to mathematicians. With this aim in view, the present paper is concerned with correlation among sums having common elements, extending and generalizing the preceding papers in several ways.

We shall assume our drawings made from a continuous universe characterized by a rather arbitrary law of distribution. We shall define n sums, each of an arbitrary number of elements, formed in such a manner that any two consecutive sums have elements in common, and inquire into the correlation between any two of these sums. The equations of the correlation surfaces will be expressed in terms of iterated integrals, the regression of each variable on the other will be shown to be linear, and the equations of the regression lines will be obtained. The coefficient of correlation may then be computed from the slopes of these lines.

Throughout this paper we shall understand a probability function, $f(t)$, to be, for all values of t on a range R, a single-valued, real-valued, non-negative, continuous function of t . It is then Riemann integrable on R , and we shall require that $\int_R f(t)dt = 1$. We define the probability that a value of t , drawn at random from the range R , lie in the interval (a, b). a and b in R and $b > a$, to be $\int_a^b f(t)dt$. We may then say that $f(t)dt$ is, to within infinitesimals of higher order, the probability that a value of t drawn at random lies in the interval $(t, t + \Delta t)$. Bachelier[5] has classified probabilities into those of the first, second, and third kinds, and Craig[6] has extended this to probability functions, according as R is the range $(-\infty, \infty)$, $(0, \infty)$, and $(0, a)$, respectively. We shall find it convenient to adopt this classification.

[5] L. Bachelier, "Calcul des Probabilities." (1912), p. 155.

[6] Allen T. Craig, "On the Distribution ^f Certain Statistics," American Journal of Mathematics, Vol. 54 (1932), pp. 353-366.

I. Sums of elements drawn from a universe characterized by a probability function of the first kind.

1. The correlation between two sums having random elements in common. Let $f(t)$, a probability function of the first kind, characterize the distribution of the variable t . Let the principal variable x_1 be defined as the sum of n_1 independent values of t drawn at random. Further, let the principal variable x_2 be defined as the sum of k_{12} random values of the n_1 values of t composing x_1 and of $n_2 - k_{12}$ independent random values of t taken directly from the universe characterized by $f(t)$.

Theorem I. *Given the sums x_1 and x_2 as defined above, with k_{12} random elements in common.*

a) *The marginal distributions of x_1 and x_2 are given, respectively, by*

$$(1.11)\quad G_1(x_1)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(t_{11})f(t_{12})\cdots f(t_{1,n_1-1})f(x_1-t_{11}-\cdots-t_{1,n_1-1})dt_{1,n_1-1}\cdots dt_{11},$$

and

$$(1.12)\quad \begin{aligned} &G_2(x_2)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(t_{11})\cdots f(t_{1k_{12}})f(t_{2,k_{12}+1})\cdots f(t_{2,n_2-1}) \\ &\times f(x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-1})dt_{2,n_2-1}\cdots dt_{2,k_{12}+1}dt_{1k_{12}}\cdots dt_{11}. \end{aligned}$$

b) *The correlation surface, $w = F(x_1, x_2)$, or the simultaneous law of distribution of x_1 and x_2, is given by*

$$(1.2)\quad \begin{aligned} &F(x_1,x_2)=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(t_{11})\cdots f(t_{1,n_1-1})f(x_1-t_{11}-\cdots-t_{1,n_1-1})f(t_{2,k_{12}+1})\cdots f(t_{2,n_2-1}) \\ &\times f(x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-1})dt_{2,n_2-1}\cdots dt_{2,k_{12}+1}dt_{1k_{12}}\cdots dt_{11}. \end{aligned}$$

c) *The regression curves of x_2 on x_1 and of x_1 on x_2 are linear, and are given, respectively, by the following equations:*

$$(1.31)\quad \bar{x}_2 = \frac{k_{12}x_1}{n_1} + (n_2 - k_{12})M,$$

and

$$\bar{x}_1 = \frac{k_{12} x_2}{n_2} + (n_1 - k_{12}) M, \tag{1.32}$$

where
$$M = \int_{-\infty}^{\infty} t f(t) dt.$$

Hence, the coefficient of correlation between x_1 *and* x_2 *is*

$$r_{x_1 x_2} = \frac{k_{12}}{(n_1 n_2)^{1/2}}.$$

Proof. The proof for the expressions for the marginal distributions of x_1 and x_2 are given by Craig[7] and need not be repeated here. The correlation surface $w = F(x_1, x_2)$ is derived by a simple extension of the same method to two independent variables.

The regression curve of x_2 on x_1 is the locus of the ordinate of the centroid $\bar{x}_2$ of a section of the surface for any given x_1. Thus

$$\bar{x}_2 = \frac{\left[\int_{-\infty}^{\infty} x_2 F(x_1, x_2) dx_2\right]}{\left[\int_{-\infty}^{\infty} F(x_1, x_2) dx_2\right]}. \tag{1.4}$$

It will be convenient in what follows to use an abbreviated notation by letting

$$\theta(x_1, t_{11}, \cdots t_{1,n_1-1}) = f(t_{11}) \cdots f(t_{1,n_1-1}) f(x_1 - t_{11} - \cdots - t_{1,n_1-1}), \tag{1.5}$$

which is merely the integrand of the marginal distribution of x_1. Where no ambiguity can result, θ_{x_1} will be used in place of $\theta(x_1, t_{11}, \cdots t_{1,n_1-1})$. Then $F(x_1, x_2)$ may be written

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta(x_1, t_{11}, \cdots t_{1,n_1-1}) \prod_{j=k_{12}+1}^{n_2-1} f(t_{2j}) f(x_2 - t_{11} - \cdots - t_{1k_{12}} - t_{2,k_{12}+1} - \cdots - t_{2,n_2-1})$$
$$\times\, dt_{2,n_2-1} \cdots dt_{2,k_{12}+1}\, dt_{1,n_1-1} \cdots dt_{11}.$$

Now let $v = x_2 - t_{11} - \cdots - t_{1k_{12}} - t_{2,k_{12}+1} - \cdots - t_{2,n_2-1}$. Changing the variable

[7] Allen T. Craig, loc. cit., pp. 355-356.

from x_2 to v, (1.4) becomes

$$(1.6)\quad \bar{x}_2 = \Big\{\Big[\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_{11} + \cdots + \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_{1k_{12}} + \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_{2,k_{12}+1} + \cdots$$

$$+ \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_{2,n_2-1} + \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} v\Big]\, \theta_{x_1} \prod_{j=k_{12}+1}^{n_2-1} f(t_{2j}) f(v)\, dt_{2,n_2-i} \cdots dt_{11}\, dv\Big\}$$

$$\Big/ \Big\{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta_{x_1} \prod_{j=k_{12}+1}^{n_2-1} f(t_{2j}) f(v)\, dt_{2,n_2-1} \cdots dt_{11}\, dv\Big\}.$$

It will be noted that the terms in the numerator fall into two groups: those terms containing the factors t_{1i}, $(i = 1, 2, \cdots k_{12})$, and those terms containing the factors v or t_{2j}, $(j = k_{12}+1, k_{12}+2, \cdots n_2-1)$. Further, since the order of integration here is immaterial, the equality of the k_{12} integrals of the first group follows readily. Similarly, the equality of the $n_2 - k_{12}$ integrals of the second group follows. The expression (1.6) may then be written

$$(1.7)\quad \bar{x}_2 = \Big\{k_{12}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_{11}\, \theta_{x_1} \prod_{j=k_{12}+1}^{n_2-1} f(t_{2j}) f(v)\, dv\, dt_{2,n_2-1} \cdots dt_{11}$$

$$+ (n_2 - k_{12})\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} v\, \theta_{x_1} \prod_{j=k_{12}+1}^{n_2-1} f(t_{2j}) f(v)\, dv\, d_{2,n_2-i} \cdots dt_{11}\Big\}$$

$$\Big/ \Big\{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta_{x_1} \prod_{j=k_{12}+1}^{n_2-1} f(t_{2j}) f(v)\, dv\, dt_{2,n_2-1} \cdots dt_{11}\Big\}$$

In (1.7), it is clear that the integrations with respect to each t_{2j} may be effected immediately, making use of $\int_{-\infty}^{\infty} f(v)\,dv = 1$ In the first term of the numerator and in the denominator the variable v may likewise be integrated out. The denominator is now equal to (1.11), the marginal distribution function of x_1. In the second term of the numerator, $v \cdot f(v)$ is independent of the remaining factors, and $\int_{-\infty}^{\infty} v f(v)\,dv$ is a constant which we shall denote by M. This second term of the numerator is now equal to $(n_2 - k_{12})M$ times the marginal distribution function of x_1

Hence, we have now reduced the expression (1.7) for $\bar{x}_2$ to the following form:

$$\text{(1.8)} \qquad \bar{x}_2 = k_{12} I_{n_1} + (n_2 - k_{12}) M,$$

where $$I_{n_1} = \frac{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_{11}\,\theta(x_1, t_{11}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{11}}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta(x_1, t_{11}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{11}}.$$

To evaluate I_n, let $t_{11} = x_1 - u - t_{12} - \cdots - t_{1,n_1-1}$.

Then
$$I_{n_1} = \frac{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_1\,\theta(x_1, u, t_{12}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{12}\,du}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta(x_1, u, t_{12}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{12}\,du}$$
$$- \frac{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} u\,\theta(x_1, u, t_{12}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{12},\,du}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta(x_1, u, t_{12}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{12}\,du}$$
$$- \sum_{j=2}^{n_1-1} \frac{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} t_{1j}\,\theta(x_1, u, t_{12}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{12}\,du}{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \theta(x_1, u, t_{12}, \cdots t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{12}\,du}.$$

The first term in the above expression for I_{n_1} is equal to x_1. Each of the remaining $n_1 - 1$ terms is equal to I_{n_1}. Hence

$$I_{n_1} = x_1 - (n_1 - 1) I_{n_1},$$

and

$$I_{n_1} = \frac{x_1}{n_1}.$$

From (1.8) and (1.9), we have

$$\bar{x}_2 = \frac{k_{12} x_1}{n_1} + (n_2 - k_{12}) M.$$

In exactly the same manner, we may show that

$$\bar{x}_1 = \frac{k_{12} x_2}{n_2} + (n_1 - k_{12}) M.$$

Making use of the fact that in the case of linear regression the square of the correlation coefficient is equal to the product of the slopes of the two lines of regression, we obtain

$$r_{x_1x_2} = \frac{k_{12}}{(n_1n_2)^{\frac{1}{2}}},$$

which completes the proof of the theorem.

Corollary. If x and y are each the sum of n independent random values of a variable t from a universe characterized by $f(t)$, and have k of these values in common, the coefficient of correlation between x and y is equal to the ratio of the number of values of t held in common to the total number composing each principal variable. Thus, $r_{xy} = \frac{k}{n}$.

This corollary of Theorem I was proved by Kapteyn[8] for the special case of a normal parent distribution of the variable t.

Illustration. As a simple illustration of the application of the foregoing theorem, let us consider the case where $x_1 = t_{11} + t_{12}$, $x_2 = t_{11} + t_{22}$ with t_{11}, t_{12}, t_{22}, as independent random drawings of t from the Gaussian distribution,

$$f(t) = (2\pi)^{-\frac{1}{2}} e^{-\frac{t^2}{2}}$$

From (1.11), the marginal distribution of x_1 is

$$G_1(x_1) = (4\pi)^{-\frac{1}{2}} e^{-\frac{x_1^2}{4}}.$$

Similarly, the marginal distribution of x_2 is

$$G_2(x_2) = (4\pi)^{-\frac{1}{2}} e^{-\frac{x_2^2}{4}}$$

The correlation surface, $w = F(x_1, x_2)$, obtained by applying (1.2), is

$$F(x_1,x_2) = e \; - \frac{\frac{(x_1^2 - x_1x_2 + x_2^2)}{3}}{(2\pi\cdot 3^{\frac{1}{2}})},$$

a normal correlation surface with $r_{x_1x_2} = \frac{1}{2}$.

2. The correlation among three sums. We now proceed to extend the preceding theorem to more than two sums. Let us define a third sum, or principal variable, x_3, as the sum of k_{23}

[8] J. C. Kapteyn, loc. cit.

elements taken at random from the n_2 values of t composing x_2 plus the sum of $n_3 - k_{23}$ independent random values of t drawn from the parent population. It is apparent, then, that the marginal distributions of x_1 , x_2 , and x_3 , and the correlation surfaces $F_1(x_1, x_2)$ and $F_2(x_2, x_3)$ will be formed exactly as were those of x_1 and x_2 in Theorem I. From this theorem, we are at once in a position to write the equations of the lines of regression and the coefficients of correlation for these surfaces. The surface $w = F(x_1, x_3)$ remains to be investigated, as does the four-dimensional surface, $v = \psi(x_1, x_2, x_3)$, which may be obtained in almost the same manner.

Theorem II. *Given* $f(t)$ *and* x_1 , x_2 , x_3 , *as defined abovc. Let* θ_{x_1}. *be defined as in* (1.5). *Let*

$$\phi(x, t_{11}, \cdots t_{1,k_{23}-g}, t_{2,k_{23}-g+1}, \cdots t_{2k_{23}}, t_{3,k_{23}+1}, \cdots t_{3,n_3-1}) =$$
$$f(t_{2,k_{23}-g+1}) \cdots f(t_{2k_{23}}) f(t_{3,k_{23}+1}) \cdots f(t_{3,n_3-1})$$
$$\times f(x - t_{11} - \cdots - t_{1,k_{23}-g} - t_{2,k_{23}-g+1} - \cdots - t_{2k_{23}} - t_{3,k_{23}+1} - \cdots - t_{3,n_3-1}).$$

If $f(t)$ *is a probability function of the first kind, then the expression for the simultaneous distribution of* x_1 *and* x_3 *is*

(2.1)
$$F(x_1, x_3) = \sum_{g=0}^{k_{23}} \left\{ \binom{k_{12}}{k_{23}-g} \binom{n_2 - k_{12}}{g} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta_{x_1} \phi(x_3, t_{11} \cdots t_{1,k_{23}-g},\right.$$
$$t_{2,k_{23}-g+1}, \cdots t_{2k_{23}}, t_{3,k_{23}+1}, \cdots t_{3,n_3-1}) dt_{3,n_3-i} \cdots dt_{3,k_{23}+1}$$
$$\left. \times\, dt_{2k_{23}} \cdots dt_{2,k_{23}-g+1}\, dt_{1,n_1-1} \cdots dt_{11} \right\} \Big/ \binom{n_2}{k_{23}},$$

where by $\binom{c}{d}$ *is understood the number of combinations of* c *items taken* d *at a time.*

Proof. Let us temporarily require that $k_{12} \geq k_{23}$. We shall show later that this restriction may be removed. The probability that x_1 and x_3 as defined contain $k_{23} - g, (g = 0, 1, 2, \cdots k_{23})$, elements in common is $\binom{k_{12}}{k_{23}-g} \binom{n_2 - k_{12}}{g} \Big/ \binom{n_2}{k_{23}}$.

The probability of the occurrence of any given pair of values (x_1, x_3), that is, the probability of a point falling into a given rectangle, $(x_1, x_1+\Delta x_1, x_3, x_3+\Delta x_3)$, is the sum of the probabilities of all of the mutually exclusive ways in which it can occur. Each of the terms in (2.1) multiplied by $dx_1 dx_3$ consists of the integral, (derived by the method of Theorem I), which is the probability, to within infinitesimals of higher order, of the occurrence of a given pair, (x_1, x_3), with a specified number of values of t in common, times a coefficient which is equal to the probability of the occurrence of this specified number of values of t in common. Each of the terms as a whole, then, is the probability that the given (x_1, x_3) will occur with a specified number of values of t in common. Hence, the expression (2.1), being the sum of the probabilities of all of the mutually exclusive ways in which x_1 and x_3 can fall within the desired rectangle, is the probability that this will occur. This establishes the theorem when $k_{12} \geqq k_{23}$.

If $k_{12} < k_{23}$, then the maximum number of values of t which x_1 and x_3 can have in common is k_{12}. The expression for $F(x_1, x_3)$ in this case, then, consists of the sum of all of the terms of (2.1) beginning with the term where x_1 and x_3 have k_{12} values of t in common and continuing to include the term derived from the case where they have no values of t in common. Equation (2.1), however, in its present form may be considered as a correct formal expression for the correlation surface even when $k_{12} < k_{23}$, since in this case all of the coefficients of the terms where x_1 and x_3 are to have more than k_{12} values of t in common are zero. This follows from the definition $\binom{c}{d} = 0$ if $c < d$. Thus

$$\binom{k_{12}}{k_{23}} = \binom{k_{12}}{k_{23}-1} = \cdots = \binom{k_{12}}{k_{12}+1} = 0 \text{ if } k_{12} < k_{23}.$$

Hence, we may now remove the restriction that $k_{12} \geqq k_{23}$. This establishes the theorem.

We are now in a position to write down the surface

$$v = \psi(x_1, x_2, x_3).$$

It is given by the following expression, where, by $t_{2,k_{23}-g+1}, \cdots t_{2,k_{23}}$ is meant any g values of the t_{2j} :

$$\psi(x_1, x_2, x_3) = \sum_{g=0}^{k_{23}} \left\{ \binom{k_{12}}{k_{23}-g} \binom{n_2-k_{12}}{g} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \theta_{x_1} f(t_{2,k_{12}+1}) \cdots f(t_{2,n_2-1}) \right.$$

$$\times f(x_2 - t_{11} - \cdots - t_{1k_{12}} - t_{2,k_{12}+1} - \cdots - t_{2,n_2-1}) f(t_{3,k_{23}+1}) \cdots f(t_{3,n_3-1})$$

$$\times f(x_3 - t_{11} - \cdots - t_{1,k_{23}-g} - t_{2,k_{23}-g+1} - \cdots - t_{2,k_{23}} - t_{3,k_{23}+1} - \cdots - t_{3,n_3-1})$$

$$\left. dt_{3,n_3-1} \cdots dt_{3,k_{23}+1} \, dt_{2,n_2-1} \cdots dt_{2,k_{12}+1} \, dt_{1,n_1-1} \cdots dt_{11} \right\}.$$

Theorem III. *The regression curves of x_3 on x_1 and of x_1 on x_3 for the correlation surface $w = F(x_1, x_3)$, defined in Theorem II, are linear and are given, respectively, by the following equations:*

$$(2.21) \qquad \bar{x}_3 = \frac{k_{12} k_{23} x_1}{n_1 n_2} + \frac{(n_2 n_3 - k_{12} k_{23}) M}{n_2},$$

and

$$(2.22) \qquad \bar{x}_1 = \frac{k_{12} k_{23} x_3}{n_2 n_3} + \frac{(n_1 n_2 - k_{12} k_{23}) M}{n_2},$$

where M is defined as in Theorem I. Further, the coefficient of correlation between x_1 and x_3 is

$$(2.3) \qquad r_{x_1 x_3} = \frac{k_{12} k_{23}}{n_2 (n_1 n_3)^{\frac{1}{2}}} = r_{x_1 x_2} r_{x_2 x_3}.$$

Proof. As in the proof of Theorem I, we set up the expression for the locus of the ordinate of the centroid of a section of the surface for a fixed x_1. We have

$$\bar{x}_3 = \frac{\int_{-\infty}^{\infty} x_3 F(x_1, x_3) \, dx_3}{\int_{-\infty}^{\infty} F(x_1, x_3) \, dx_3}$$

where $F(x_1, x_3)$ is given by (2.1). From the definition of a marginal distribution, we know that $\int_{-\infty}^{\infty} F(x_1, x_3)\,dx_3$ reduces to (1.11), the marginal distribution of x_1. Let us now write the expression for $\bar{x}_3$ as the sum of $k_{23}+1$ fractions. Thus

$$(2.4)\quad \bar{x}_3 = \sum_{g=0}^{k_{23}} \binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}\Big\{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_3\,\theta_{x_1}$$

$$\times\,\phi(x_3, t_{11}, \cdots t_{1,k_{23}-g}, t_{2,k_{23}-g+1}, \cdots t_{2k_{23}}, t_{3,k_{23}+1}, \cdots t_{3,n_3-1})$$

$$\times\,dt_{3,n_3-1}\cdots dt_{3,k_{23}+1}\,dt_{2k_{23}}\cdots dt_{2,k_{23}-g+1}\,dt_{1,n_1-1}\cdots dt_{11}\Big\}$$

$$\Big/\Big\{\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\theta_{x_1}\,dt_{1,n_1-1}\cdots dt_{11}\binom{n_2}{k_{23}}.$$

Hereafter, we shall call an expression of the form

$$\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}\Big/\binom{n_2}{k_{23}}$$

a "probability coefficient." Then (2.4) is the sum of products, each of which is a probability coefficient times an expression which is equivalent to the expression for $\bar{x}_3$ for the simple case where x_3 would be derived directly from x_1 by the drawing of $k_{23}-g$ values of t from x_1. These latter expressions, by the application of Theorem I, may each be written in the same form as (1.3). Hence, (2.4) has been reduced to

$$(2.5)\quad \bar{x}_3 = x_1\left[\frac{1}{n_1}\sum_{g=0}^{k_{23}-1}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}(k_{23}-g)\right]\Big/\binom{n_2}{k_{23}}$$

$$+M\left[\sum_{g=0}^{k_{23}}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}(n_3-k_{23}+g)\right]\Big/\binom{n_2}{k_{23}}$$

$$=\frac{x_1 k_{12}}{n_1}\left[\sum_{g=0}^{k_{23}-1}\binom{k_{12}-1}{k_{23}-g-1}\binom{n_2-k_{12}}{g}\right]\Big/\binom{n_2}{k_{23}}$$

$$+M\left[(n_3-k_{23})\sum_{g=0}^{k_{23}}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}\right.$$

$$\left.+\sum_{g=0}^{k_{23}}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}g\right]\Big/\binom{n_2}{k_{23}}.$$

By the use of a well-known theorem of combinatory analysis,[9] we have that

$$\frac{1}{n_1}\sum_{g=0}^{k_{23}-1}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}(k_{23}-g)\Big/\binom{n_2}{k_{23}}=\frac{k_{12}}{n_1}\binom{n_2-1}{k_{23}-1}\Big/\binom{n_2}{k_{23}}=\frac{k_{12}k_{23}}{n_1n_2},$$

and

$$(n_3-k_{23})\sum_{g=0}^{k_{23}}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}\Big/\binom{n_2}{k_{23}}=\binom{n_2}{k_{23}}(n_3-k_{23})\Big/\binom{n_2}{k_{23}}=(n_3-k_{23}).$$

Moreover,

$$\sum_{g=0}^{k_{23}}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}g\Big/\binom{n_2}{k_{23}}=(n_2-k_{12})\sum_{g=1}^{k_{23}}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}-1}{g-1}\Big/\binom{n_2}{k_{23}},$$

which reduces to

$$(n_2-k_{12})\binom{n_2-1}{k_{23}-1}\Big/\binom{n_2}{k_{23}}=(n_2-k_{12})k_{23}\big/n_2$$

by the same theorem of combinatory analysis.

Hence, (2.5) becomes

$$\bar{x}_3=\frac{k_{12}\,k_{23}\,x_1}{n_1n_2}+\frac{(n_2\,n_3-k_{12}\,k_{23})M}{n_2}.$$

In exactly the same manner, we may show that

$$\bar{x}_1=\frac{k_{12}\,k_{23}\,x_3}{n_2\,n_3}+\frac{(n_1\,n_2-k_{12}\,k_{23})M}{n_2}.$$

We then obtain the coefficient of correlation from the slopes of these lines. It is

$$r_{x_1x_3}=\frac{k_{12}\,k_{23}}{n_2(n_1\,n_3)^{1/2}}=r_{x_1x_2}\,r_{x_2x_3}.$$

This completes the proof of the theorem, since

$$r_{x_1x_2}=\frac{k_{12}}{(n_1\,n_2)^{1/2}}\qquad r_{x_2x_3}=\frac{k_{23}}{(n_2\,n_3)^{1/2}}.$$

3. The correlation among p sums. We now extend our discussion to p principal variables, forming each successive one

[9] E. Netto, "Lehrbuch der Combinatorik," (1901), pp. 12-13.

in the same manner in which x_2 and x_3 were formed above; that is, x_i, $(i = 2, 3, \cdots p)$, is equal to the sum of $k_{i-1,\, i}$ random drawings of t from the constituent values of t forming x_{i-1}, plus the sum of $n_i - k_{i-1,\, i}$ independent random drawings of t directly from the universe characterized by $f(t)$. The correlation surface, $w = F(x_1, x_p)$, can at once be written in the same manner as the surface considered in Theorem II. That is, each term of the expression for $F(x_1, x_p)$, multiplied by $dx_1 . dx_p$, consists of an iterated integral which represents the probability, to within infinitesimals of higher order, of the occurrence of a given pair, (x_1, x_p), with a specified number of values of t in common, times a probability coefficient which represents the probability of the occurrence of this specified number of values of t in common. This same method may be employed in writing the correlation surface for any pair of principal variables. The expressions for the probability coefficients, however, become increasingly complex as the number of ways in which the two principal variables can have $0, 1, 2, \cdots$ values of t in common increases.

The following theorem can be proved by mathematical induction. The proof is not difficult, though tedious, and on that account will not be presented here.

Theorem IV. *If $f(t)$ is a probability function of the first kind, and $F(x_1, x_p)$ is the simultaneous law of distribution of x_1 and x_p, then the regression of x_1 on x_p and of x_p on x_1 are linear and are given, respectively, by the following equations:*

$$(3.1) \quad \bar{x}_p = \frac{k_{12}\, k_{23} \cdots k_{p-1,\, p}}{n_1\, n_2 \cdots n_{p-1}}\, x_1 + \frac{n_2\, n_3 \cdots n_p - k_{12}\, k_{23} \cdots k_{p-1,\, p}}{n_2\, n_3 \cdots n_{p-1}}\, M,$$

$$(3.2) \quad \bar{x}_1 = \frac{k_{12}\, k_{23} \cdots k_{p-1,\, p}}{n_2\, n_3 \cdots n_p}\, x_p + \frac{n_1\, n_2 \cdots n_{p-1} - k_{12}\, k_{23} \cdots k_{p-1,\, p}}{n_2\, n_3 \cdots n_{p-1}}\, M.$$

Further, the coefficient of correlation between x_1 and x_p is

$$(3.3) \quad r_{x_1 x_p} = \frac{k_{12}\, k_{23} \cdots k_{p-1,\, p}}{n_2\, n_3 \cdots n_{p-1} (n_1 n_p)^{\frac{1}{2}}} = r_{x_1 x_2} \cdot r_{x_2 x_3} \cdots r_{x_{p-1} x_p}.$$

II. Sums of elements drawn from a universe characterized by a probability function of the second kind.

4. The correlations between two sums. Let $f(t)$, a probability function of the second kind, characterize the distribution of the variable t. Let the principal variable x_1 be defined as the sum of n_1 independent values of t drawn at random. Further, let the principal variable x_2 be defined as the sum of k_{12} random values of the n_1 values of t composing x_1 and of n_2-k_{12} independent random values of t taken directly from the universe characterized by $f(t)$.

Theorem V. *Given the sums x_1 and x_2 as defined above with k_{12} random elements in common.*

a) *The marginal distributions of x_1 and x_3 are given, respectively, by*

$$(4.11)\quad G_1(x_1)=\int_0^{x_1}\int_0^{x_1-t_{11}}\cdots\int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}} f(t_{11})\cdots f(t_{1,n_1-1})$$

$$\times f(x_1-t_{11}-\cdots-t_{1,n_1-1})\,dt_{1,n_1-1}\cdots dt_{11},$$

and

$$(4.12)\quad G_2(x_2)=\int_0^{x_2}\int_0^{x_2-t_{11}}\cdots\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-2}}$$

$$\times f(t_{11})\cdots f(t_{1k_{12}})f(t_{2,k_{12}+1})\cdots f(t_{2,n_2-1})$$

$$\times f(x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-1})\,dt_{2,n_2-1}\cdots dt_{2,k_{12}+1}\,dt_{1k_{12}}\cdots dt_{11}.$$

b) *The correlation surface, $w=F(x_1,x_2)$, which is in two distinct parts joined along the plane $x_1 - x_2 = 0$, is given by*

$$(4.2a)\quad F_1(x_1,x_2)=\int_0^{x_2}\int_0^{x_2-t_{11}}\cdots\int_0^{x_2-t_{11}-\cdots-t_{1,k_{12}-1}}\int_0^{x_1-t_{11}-\cdots-t_{1,k_{12}-1}-t_{1k_{12}}}$$

$$\int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}}\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}}\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}}$$

$$\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-2}} \theta(x_1,t_{11},\cdots t_{1,n_1-1})$$

$$\times \phi(x_2, t_{11}, \cdots t_{1k_{12}}, t_{2,k_{12}+1}, \cdots t_{2,n_2-1}) dt_{2,n_2-1} \cdots$$

$$\times \, dt_{2,k_{12}+1} \, dt_{1,n_1-1} \cdots dt_{11}.$$

$$(x_2 \leqq x_1 < \infty);$$

(4.2b)

$$F_2(x_1, x_2) = \int_0^{x_1} \int_0^{x_1 - t_{11}} \cdots \int_0^{x_1 - t_{11} - \cdots - t_{1,n_1-2}} \int_0^{x_2 - t_{11} - \cdots - t_{1k_{12}}}$$

$$\cdot \int_0^{x_2 - t_{11} - \cdots - t_{1k_{12}} - t_{2,k_{12}+1}} \cdots \int_0^{x_2 - t_{11} - \cdots - t_{1k_{12}} - t_{2,k_{12}+1} - \cdots - t_{2,n_2-2}}$$

$$\times \theta(x_1, t_{11}, \cdots t_{1,n_1-1}) \phi(x_2, t_{11}, \cdots t_{1k_{12}}, t_{2,k_{12}+1}, \cdots t_{2,n_2-1})$$

$$\times \, dt_{2,n_2-1} \cdots dt_{2,k_{12}+1} \, dt_{1,n_1-1} \cdots dt_{11}.$$

$$(x_1 \leqq x_2 < \infty).$$

c) *The regression curves of* x_2 *on* x_1 *and of* x_1 *on* x_2 *are linear and are given, respectively, by the following equations:*

(1.31)
$$\bar{x}_2 = \frac{k_{12} x_1}{n_1} + (n_2 - k_{12}) M,$$

and

(1.32)
$$\bar{x}_1 = \frac{k_{12} x_2}{n_2} + (n_1 - k_{12}) M,$$

where
$$M = \int_0^\infty t \, f(t) \, dt.$$

Hence, the cofficient of correlation between x_1 *and* x_2 *is*

$$r_{x_1 x_2} = \frac{k_{12}}{(n_1 n_2)^{\frac{1}{2}}}.$$

Proof. The proof for the marginal distributions of x_1 and of x_2 are given by Craig[10] and need not be repeated here. The expressions tor the correlation surface are derived by a simple extension of the same method to two independent variables. The

[10] Allen T. Craig, loc. cit., p. 356.

limits of integration may be easily verified.

As in the proof of Theorem I, the regression of x_2 on x_1 is given by the locus of the ordinate of the centroid of the section of the surface for a given x_1. However, as the surface here is in two distinct, but connected, parts, we have two terms in both numerator and denominator. The expression for $\bar{x}_2$ is

$$(4.3) \qquad x_2 = \frac{\int_0^{x_1} x_2 F_1'(x_1, x_2)\,dx_2 + \int_{x_1}^{\infty} x_2 F_2'(x_1, x_2)\,dx_2}{\int_0^{x_1} F_1'(x_1, x_2)\,dx_2 + \int_{x_1}^{\infty} F_2'(x_1, x_2)\,dx_2},$$

where $F_1'(x_1, x_2)$ and $F_2'(x_1, x_2)$ are defined by (4.2a) and (4.2b), respectively.

In the paragraphs immediately following, we shall be concerned principally with interchanging the order of integration, with the accompanying changes in the limits. It will be convenient to write the differential immediately following its respective integral sign. Consider the first term of the numerator. Successive interchanging the order of integration between integration with respect to x_2 and with respect to $t_{11}, t_{12}, \ldots t_{1k_{12}}$, respectively, and making the appropriate changes in the limits, we get, writing Φ_{x_2} for $\Phi(x_2, t_{11}, \cdots t_{1k_{12}}, t_{2,k_{12}+1}, \cdots t_{2,n_2-1})$

$$(4.4) \int_0^{x_1} dt_{11} \int_0^{x_1-t_{11}} dt_{12} \cdots \int_0^{x_1-t_{11}-t_{12}-\cdots-t_{1,k_{12}-1}} dt_{1k_{12}} \int_{t_{11}+t_{12}+\cdots+t_{1k_{12}}}^{x_1} dx_2$$
$$\int_0^{x_1-t_{11}-\cdots-t_{1k_{12}}} dt_{1,k_{12}+1} \cdots \int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}} dt_{1,n_1-1} \int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}} dt_{2,k_{12}+1}$$
$$\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}} dt_{2,k_{12}+2} \cdots \int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-2}} dt_{2,n_2-1}$$
$$\times\ x_2\, \theta_{x_1}\, \Phi_{x_2}.$$

Now consider the second term of the numerator of (4.3). As the limits are constants with respect to the variables of integration

$x_2, t_{11}, \ldots . t_{1k_{12}}$, we may interchange the order of integration successively until we have

(4.5)

$$\int_0^{x_1} dt_{11} \int_0^{x_1-t_{11}} dt_{12} \cdots \int_0^{x_1-t_{11}-\cdots-t_{1,k_{12}-1}} dt_{1k_{12}} \int_{x_1}^{\infty} dx_2$$
$$\int_0^{x_1-t_{11}-\cdots-t_{1k_{12}}} dt_{1,k_{12}+1} \cdots \int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}} dt_{1,n_1-1}$$
$$\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}} dt_{2,k_{12}+1} \int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}} dt_{2,k_{12}+2} \cdots$$
$$\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-2}} dt_{2,n_2-1}\, x_2\, \theta_{x_1}\, \phi_{x_2} .$$

We may now combine the first and second terms, (4.4) and (4.5), getting

$$\int_0^{x_1} dt_{11} \int_0^{x_1-t_{11}} dt_{12} \cdots \int_0^{x_1-t_{11}-\cdots-t_{1,k_{12}-1}} dt_{1k_{12}} \int_{t_{11}+t_{12}+\cdots+t_{1k_{12}}}^{\infty} dx_2$$
$$\int_0^{x_1-t_{11}-\cdots-t_{1k_{12}}} dt_{1,k_{12}+1} \cdots \int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}} dt_{1,n_1-1}$$
$$\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}} dt_{2,k_{12}+1} \int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}} dt_{2,k_{12}+2} \cdots$$
$$\int_0^{x_2-t_{11}-\cdots-t_{1k_{12}}-t_{2,k_{12}+1}-\cdots-t_{2,n_2-2}} dt_{2,n_2-1}\, x_2\, \theta_{x_1}\, \phi_{x_2} .$$

As the limits of integration are constant with respect to the variables x_2 and $t_{1,k_{12}+1}, \ldots t_{1,n_1-1}$, we may at once interchange successively the orders of integration with respect to x_2 and with respect to $t_{2,k_{12}+1}, t_{2,k_{12}+2}, \ldots\ldots . t_{2,n_2-1}$, respectively, making the proper changes in the limits. We then have

(4.6)

$$\int_0^{x_1} dt_{11} \cdots \int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}} dt_{1,n_1-1} \int_0^{\infty} dt_{2,k_{12}+1} \cdots$$
$$\int_0^{\infty} dt_{2,n_2-1} \int_{t_{11}+\cdots+t_{1k_{12}}+t_{2,k_{12}+1}+\cdots+t_{2,n_2-1}}^{\infty} dx_2\ x_2\, \theta_{x_1}\, \phi_{x_2} .$$

The denominator of (4.3) may be reduced to this same form except for the absence of the factor x_2 in the integrand.

Let us make the transformation

$$v = x_2 - t_{11} - \cdots - t_{1k_{12}} - t_{2,k_{12}+1} - \cdots - t_{2,n_2-1},$$

as was done in the proof of Theorem I. The limits $t_{11} + \cdots + t_{2n_2-1}$ to ∞ on x_2 now become 0 to ∞ on v. We have now reduced (4.3) to the following form:

(4.7)

$$\bar{x}_2 = \Bigg\{\Bigg[\int_0^{x_1}\cdots\int_0^{\infty} t_{11} + \int_0^{x_1}\cdots\int_0^{\infty} t_{12} + \cdots\cdots$$

$$\int_0^{x_1}\cdots\int_0^{\infty} t_{1k_{12}} + \int_0^{x_1}\cdots\int_0^{\infty} t_{2,k_{12}+1} + \cdots + \int_0^{x_1}\cdots\int_0^{\infty} t_{2,n_2-1}$$

$$+ \int_0^{x_1}\cdots\int_0^{\infty} v\Bigg]\theta_{x_1}\prod_{j=k_{12}+1}^{n_2-1} f(t_{2j})f(v)\,dv\,dt_{2,n_2-1}\cdots$$

$$dt_{2,k_{12}+1}\,dt_{1,n_1-1}\cdots dt_{11}\Bigg\}\Bigg/$$

$$\Bigg\{\int_0^{x_1}\cdots\int_0^{\infty}\theta_{x_1}\prod_{j=k_{12}+1}^{n_2-1} f(t_{2j})f(v)\,dv\,dt_{2,n_2-1}\cdots dt_{2,k_{12}+1}\,dt_{1,n_1-1}\cdots dt_{11}\Bigg\}.$$

The denominator reduces at once to $G(x_1)$ in (4.11). As in the proof of Theorem I directly following equation (1.6), it will be noted that the terms of the numerator fall into two groups: those k_{12} terms containing the factor t_{1i}, $(i=1,2,\ldots\ldots k_{12})$, and the n_2-k_{12} terms containing the factor v or t_{2j}, $(j=k_{12}+1,\ldots\ldots n_2-1)$. As the limits of integration with respect to each of these letter variables are 0 and ∞, and since complete interchangeability of the order of integration is then permissible, it is readily seen that any two of these n_2-k_{12} terms are equivalent. The sum of the entire group, then, may be written

(4.8)

$$(n_2-k_{12})\int_0^{x_1} dt_{11}\cdots\int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}} dt_{1,n_1-1}\int_0^{\infty} dt_{2,k_{12}+1}\cdots$$

$$\int_0^{\infty} dt_{2,n_2-1}\int_0^{\infty} dv\; v\theta_{x_1}\prod_{j=k_{12}+1}^{n_2-1} f(t_{2j})\,f(v).$$

In (4.8), it is clear that the integrations with respect to each t_{2j} may be effected immediately by making use of the hypothesis that $\int_0^\infty f(t)dt = 1$. This leaves $v f(v)\theta_{x_1}$ remaining as the integrand. The $\int_0^\infty v f(v) dv$ is a constant which we shall designate by M. Removing this constant from under the integral signs leaves us merely tne expression for the marginal distribution of x_1 times $M(n_2-k_{12})$. We then have

$$(4.9)\quad \bar{x}_2 = (n_2-k_{12})M + \sum_{i=1}^{k_{12}} \frac{\int_0^{x_1} dt_{11} \cdots \int_0^\infty dv\, t_{1i}\, \theta_{x_1} \prod_{j=k_{12}+1}^{n_2-1} f(t_{2j}) f(v)}{\int_0^{x_1} dt_{11} \cdots \int_0^{x_1 - t_{11} - \cdots - t_{1,n_1-2}} dt_{1,n_1-1}\, \theta_{x_1}}.$$

That each term in the summation in the right member of (4.9) is equal to any other term in the summation, follows from the complete interchangeability of the order of integration of any two consecutive variables, provided a corresponding interchange between these two variables is likewise carried out in the limits of integration. By successive interchanges of variables we may put the original t_{11}, t_{12}, ... $t_{1k_{12}}$ in any order we choose. Hence, the sum of the last k_{12} terms of (4.9) may be written as k_{12} times any one of them. For definiteness, select the one containing the factor t_{11} in the integrand of the numerator. We may now integrate out all of the t_{2j}, $(j = k_{12}+1, \ldots n_2-1)$ and the v exactly as before. Equation (4.9) then becomes

$$\bar{x}_2 = (n_2-k_{12})M + k_{12} \frac{\int_0^{x_1} dt_{12} \cdots \int_0^{x_1 - t_{12} - \cdots - t_{1,n_1-1}} dt_{11}\, t_{11}\, \theta_{x_1}}{\int_0^{x_1} dt_{12} \cdots \int_0^{x_1 - t_{12} - \cdots - t_{1,n_1-1}} dt_{11}\, \theta_{x_1}},$$

or $\bar{x}_2 = (n_2-k_{12})M + k_{12} I_{n_1}$.

It is not difficult to show that $I_{n_1} = \frac{x_1}{n_1}$. Hence, we have

$$\bar{x}_2 = \frac{k_{12}x_1}{n_1} + (n_2-k_{12})M.$$

In exactly the same manner, we may show that

$$\bar{x}_1 = \frac{k_{12}x_2}{n_2} + (n_1-k_{12})M.$$

The coefficient of correlation between x_1 and x_2 is

$$r_{x_1 x_2} = \frac{k_{12}}{(n_1 n_2)^{1/2}} \, ,$$

which completes the proof of the theorem.

Illustration. Consider the two sums, $x_1 = t_{11} + t_{12}$, and $x_2 = t_{11} + t_{22}$, with t_{11}, t_{12}, t_{22}, as random drawings of t from the distribution characterized by the function $f(t) = e^{-t}$ for t on the range 0 to ∞. From (4.11), the marginal distribution of x_1 is

$$G_1(x_1) = x_1 e^{-x_1}.$$

Similarly, the marginal distribution of x_2 is

$$G_2(x_2) = x_2 e^{-x_2}.$$

The correlation surface, obtained by applying (4.2a) and (4.2b), is

$$F_1(x_1, x_2) = e^{-x_1}(1 - e^{-x_2}), \quad (0 \leq x_2 \leq x_1);$$

and

$$F_2(x_1, x_2) = e^{-x_2}(1 - e^{-x_1}), \quad (x_1 \leq x_2 < \infty).$$

5. The correlation among more than two sums. We shall state, without proof, the following theorems.

Theorem VI. *Given a probability function,* $f(t)$, *of the second kind, and three principal variables,* x_1, x_2, x_3, *defined as for, Theorem II. Then the correlation surface* $w = F(x_1, x_3)$ *is given by*

(5.1a)

$$F_1(x_1, x_3) = \frac{1}{\binom{n_2}{k_{23}}} \sum_{g=0}^{k_{23}} \binom{k_{12}}{k_{23}-g} \binom{n_2 - k_{12}}{g} \int_0^{x_3} dt_{11} \cdots \int_0^{x_3 - t_{11} - \cdots - t_{1,k_{23}-g-1}} dt_{1,k_{23}-g}$$

$$\int_0^{x_1 - t_{11} - \cdots - t_{1,k_{23}-g}} dt_{1,k_{23}-g+1} \cdots \int_0^{x_1 - t_{11} - \cdots - t_{1,n_1-2}} dt_{1,n_1-1}$$

$$\int_0^{x_3 - t_{11} - \cdots - t_{1,k_{23}-g}} dt_{2,k_{23}-g+1} \cdots$$

$$\int_0^{x_3 - t_{11} - \cdots - t_{1,k_{23}-g} - t_{2,k_{23}-g+1} - \cdots - t_{2,k_{23}-1}} dt_{2 k_{23}}$$

$$\int_0^{x_3-t_{11}-\cdots-t_{1,k_{23}-g}-t_{2,k_{23}-g+1}-\cdots-t_{2k_{23}}} dt_{3,k_{23}+1}\cdots$$

$$\int_0^{x_3-t_{11}-\cdots-t_{3,n_3-2}} dt_{3,n_3-1}\,\theta_{x_1}\,\phi(x_3, t_{11}, \cdots t_{1,k_{23}-g},$$

$$t_{2,k_{23}-g+1}, \cdots t_{2k_{23}}, t_{3,k_{23}+1}, \cdots t_{3,n_3-1}),$$

$$(x_3 \leqq x_1 < \infty);$$

and

(5.1b)

$$F_2(x_1,x_3)=\frac{1}{\binom{n_2}{k_{23}}}\sum_{g=0}^{k_{23}}\binom{k_{12}}{k_{23}-g}\binom{n_2-k_{12}}{g}\int_0^{x_1} dt_{11}\cdots\int_0^{x_1-t_{11}-\cdots-t_{1,n_1-2}} dt_{1,n_1-1}$$

$$\int_0^{x_3-t_{11}-\cdots-t_{1,k_{23}-g}} dt_{2,k_{23}-g+1}\cdots$$

$$\int_0^{x_3-t_{11}-\cdots-t_{1,k_{23}-g}-t_{2,k_{23}-g+1}-\cdots-t_{2,k_{23}-1}} dt_{2k_{23}}$$

$$\int_0^{x_3-t_{11}-\cdots-t_{2,k_{23}-1}-t_{2,k_{23}}} dt_{3,k_{23}+1}\cdots$$

$$\int_0^{x_3-t_{11}-\cdots-t_{1,k_{23}-g}-t_{2,k_{23}-g+1}-\cdots-t_{2k_{23}}-t_{3,k_{23}+1}-\cdots-t_{3,n_3-2}} dt_{3,n_3-1}$$

$$\theta_{x_1}\,\phi(x_3, t_{11}, \cdots t_{1,k_{23}-g}, t_{2,k_{23}-g+1}, \cdots t_{2k_{23}}, t_{3,k_{23}+1}, \cdots t_{3,n_3-1})$$

$$(x_1 \leqq x_3 < \infty).$$

Theorem VII. *The regression curves of x_3 on x_1 and of x_1 on x_3 of the correlation surface in Theorem VI are linear and are given, respectively, by the following equations:*

(2.21)
$$\bar{x}_3 = \frac{k_{12}\,k_{23}\,x_1}{n_1 n_2} + \frac{(n_2 n_3 - k_{12}k_{23})M}{n_2},$$

and

(2.22)
$$\bar{x}_1 = \frac{k_{12}\,k_{23}\,x_3}{n_2 n_3} + \frac{(n_1 n_2 - k_{12}k_{23})M}{n_2},$$

where M is defined as in Theorem V. Further, the coefficient of correlation between x_1 and x_3 is

$$r_{x_1 x_3} = \frac{k_{12} k_{23}}{n_2 (n_1 n_3)^{1/2}} = r_{x_1 x_2} r_{x_2 x_3} . \tag{2.3}$$

Theorem VIII. The statement of this theorem differs from that of Theorem IV only in that $f(t)$ is now to be a probability function of the second kind.

III. Sums of elements drawn from a universe characterized by a probability function of the third kind.

6. The correlation between two sums. We shall now consider principal variables defined as the sums of values of t drawn from a universe characterized by $f(t)$, a probability function of the third kind, defined on the range 0 to a, and with

$$\int_0^a f(t)\,dt = 1.$$

The correlation surfaces are not developed with the same degree of generality as were those in the preceding pages because of the tediousness of the labor involved and the complexity of the correlation surface, which may consist of many sections joined together. Thus, if x is the sum of m values of t and y the sum of n, all drawn from a universe characterized by a probability function of the third kind, the correlation surface, $w = F(x,y)$, consists of $2(mn-1)$ sections, each having its own equation. Hence, only the case where x and y each consist of the sum of two values of t, with one of these held in common, will be considered here.

Theorem IX. *Let $f(t)$, a probability function of the third kind, characterize the distribution of a variable t. Let the principal variables x and y be defined by the relations $x = t_{11} + t_{12}$, $y = t_{11} + t_{22}$, where t_{11}, t_{12}, t_{22}, are independent random drawings of t from the universe.*

a.) *The marginal distributions of x and of y are given by*

(6.11)

$$G_1(x) = \int_0^x f(t)f(x-t)dt, \qquad (0 \leqq x \leqq a);$$
$$= \int_{x-a}^a f(t)f(x-t)dt, \qquad (a \leqq x \leqq 2a);$$

and

(6.12)

$$G_2(y) = \int_0^y f(t)f(y-t)dt, \qquad (0 \leqq y \leqq a);$$
$$= \int_{y-a}^a f(t)f(y-t)dt, \qquad (a \leqq y \leqq 2a).$$

b) *The correlation surface,* $w = F(x,y)$ *, is given by*

(6.2)

$$F(x,y) = \int_0^y f(t)f(x-t)f(y-t)dt, \quad (0 \leqq y \leqq x \leqq a);$$
$$= \int_0^x f(t)f(x-t)f(y-t)dt, \quad (0 \leqq x \leqq y \leqq a);$$
$$= \int_{y-a}^x f(t)f(x-t)f(y-t)dt, \quad (a \leqq y \leqq x+a \leqq 2a);$$
$$= \int_{x-a}^y f(t)f(x-t)f(y-t)dt, \quad (0 \leqq x-a \leqq y \leqq a);$$
$$= \int_{x-a}^a f(t)f(x-t)f(y-t)dt, \quad (a \leqq y \leqq x \leqq 2a);$$
$$= \int_{y-a}^a f(t)f(x-t)f(y-t)dt, \quad (a \leqq x \leqq y \leqq 2a).$$

In a) and b) above, the subscripts have been omitted from the t_{11} .

c) *The regression curves of* y *on* x *and of* x *on* y *are linear and are given, respectively, by the following equations:*

(6.31) $$\bar{y} = \frac{x}{2} + M,$$

(6.32) *and* $$\bar{x} = \frac{y}{2} + M,$$

where $$M = \int_0^a t f(t)dt.$$

Hence, the coefficient of correlation between x *and* y *is* $\frac{1}{2}$.

This theorem is a direct generalization of Rietz's paper in Biometrika cited in the introduction to this paper. The proof may be supplied by the reader.

Illustration. Let us consider the rectangular distribution given by $f(t)=\frac{1}{a}$, for t on the range 0 to a, and a to 0. This is the parent distribution in Rietz's case when $a=1$. From (6.11), the marginal distribution of x is

$$
\begin{aligned}
G_1(x) &= \frac{x}{a^2}, && (0 \leqq x \leqq a);\\
&= \frac{(2a-x)}{a^2}, && (a \leqq x \leqq 2a).
\end{aligned}
$$

Similarly, the marginal distribution of y is

$$
\begin{aligned}
G_2(y) &= \frac{y}{a^2}, && (0 \leqq y \leqq a);\\
&= \frac{(2a-y)}{a^2}, && (a \leqq y \leqq 2a).
\end{aligned}
$$

The application of (6.2) yields

$$
\begin{aligned}
F(x,y) &= \frac{y}{a^3}, && (0 \leqq y \leqq x \leqq a);\\
&= \frac{x}{a^3}, && (0 \leqq x \leqq y \leqq a);\\
&= \frac{(x-y+a)}{a^3}, && (a \leqq y \leqq x+a \leqq 2a);\\
&= \frac{(y-x+a)}{a^3}, && (0 \leqq x-a \leqq y \leqq a);\\
&= \frac{(2a-x)}{a^3}, && (a \leqq y \leqq x \leqq 2a);\\
&= \frac{(2a-y)}{a^3}, && (a \leqq x \leqq y \leqq 2a).
\end{aligned}
$$

These results, obtained directly by the use of Theorem IX, agree with those obtained by Rietz in the above-mentioned paper.

Carl H. Fischer.

ON THE CORRELATION BETWEEN CERTAIN AVERAGES FROM SMALL SAMPLES*

By

ALLEN T. CRAIG

1. *Introduction.* It is well known that no correlation exists between the arithmetic mean and standard deviation of samples drawn at random from a normal universe. However, there seems to be in the literature no treatment of the correlation between other averages either for normal or non-normal universes. In the present paper, a few simple theorems are established which make possible the determination of the type of regression of the median on the arithmetic mean, of the range on the median, and of the range on the arithmetic mean. In case the regression is linear, the coefficient of correlation may be computed.

We shall understand a probability function $f(x)$ of a real variable x to be, for all values of x on a range of R a single-valued, non-negative, continuous function with $\int_R f(x)dx = 1.$ Then $\int_a^b f(x)dx$ is the probability that a value of x chosen at random lies in the interval (a, b) where a and b are in R and $a < b$; and $f(x)dx$ is, to within infinitesimals of higher order, the probability that a value of x chosen at random lies in the interval $(x, x+dx)$. It will prove convenient to classify probability functions according as R is the range $(-\infty, \infty)$, $(0, \infty)$, or $(0, k)$, $k>0$. In accord with this classification,[1] we shall refer to probability functions as of the first, second, and third kinds respectively. In a similar manner, we define a probability function $F(x,y)$ of two independent variables.

*Presented to the American Mathematical Society, Dec. 28, 1931.

[1] Cf. L. Bachelier, Calcul des Probabilités, p. 155.

2. The correlation between the arithmetic mean $\bar{x}$ and the range W.

Theorem I. *Let $f(x)$ be the probability function of the variable x. Let $F_1(\bar{x}, W)$ be that of the arithmetic mean $\bar{x}$ and the range W in samples of three independent values of x. If $f(x)$ is a probability function of the first kind, then*

$$F_1(\bar{x}, W) = 18\int_{\bar{x}+\frac{W}{3}}^{\bar{x}+\frac{2W}{3}} f(x_1)\, f(x_1 - W)\, f(3\bar{x} - 2x_1 + W)\, dx_1 .$$

Proof. Let x_1, x_2, x_3, be the three observed values of x. Write

$$\begin{aligned} x_1 + x_2 + x_3 &= 3\bar{x}, \\ x_1 \qquad - x_3 &= W, \\ x_3 \le x_2 &\le x_1 . \end{aligned}$$

For $\bar{x}$ assigned, $-\infty < \bar{x} < \infty$, and W assigned, $0 \le W < \infty$ we must have

$$\begin{aligned} \bar{x} + \frac{W}{3} \le x_1 &\le \bar{x} + \frac{2W}{3}, \\ x_3 &= x_1 - W, \\ x_2 &= 3\bar{x} - x_1 - x_3 . \end{aligned}$$

If we consider all possible arrangements of x_1, x_2, x_3, we have

$$F_1(\bar{x}, W)\, d\bar{x}\, dW = 6\int_{\bar{x}+\frac{W}{3}}^{\bar{x}+\frac{2W}{3}} f(x_1)\, f(x_2)\, f(x_3)\, dx_1\, dx_2\, dx_3 .$$

Let

$$\begin{aligned} x_1 &= x_1 \\ x_2 &= 3\bar{x} - x_1 - x_3 , \\ x_3 &= x_1 - W. \end{aligned}$$

The absolute value of the Jacobin is 3. Hence the theorem.

In the case of samples of four independent items x_1, x_2, x_3, x_4, the probability function $F_1(\bar{x},W)$ is given by

$$F_1(\bar{x},W) = 48\int_{\bar{x}+\frac{W}{4}}^{\bar{x}+\frac{W}{2}}\int_{4\bar{x}-3x_1+W}^{x_1} f(x_1)f(x_2)f(4\bar{x}-2x_1-x_2+W)f(x_1-W)\,dx_2\,dx_1$$

$$+\,48\int_{\bar{x}+\frac{W}{2}}^{\bar{x}+\frac{3W}{4}}\int_{x_1-W}^{4\bar{x}-3x_1+2W} f(x_1)f(x_2)f(4\bar{x}-2x_1-x_2+W)f(x_1-W)\,dx_2\,dx_1.$$

We note that the probability function is made up of the sum of two parts depending on whether x_1 is in the interval $(\bar{x}+\frac{W}{4},\ \bar{x}+\frac{W}{2})$ or in the interval $(\bar{x}+\frac{W}{2},\bar{x}+\frac{3W}{4})$. Moreover, it may be of interest to note the overlapping of the ranges of integration of x_2 . To prove that $F_1(\bar{x},W)$ is given as stated, we take

$$\begin{aligned} &x_1+x_2+x_3+x_4 = 4\bar{x},\\ (1)\qquad &x_4 \le x_3,\ x_2 \le x_1,\\ &x_1 - x_4 = W. \end{aligned}$$

From (1) it readily follows that

$$(2)\qquad 2x_1 + x_2 + x_3 = 4\bar{x} + W.$$

For assigned values of $\bar{x}$ and W, the upper limit on x_1 is found from (2) by taking $x_2 = x_3 = x_4 = x_1 - W$. Thus $x_1 = \bar{x} + \frac{3W}{4}$. Similarly, the lower limit on x_1 is found from (2) by taking $x_2 = x_3 = x_1$. Thus $x_1 = \bar{x} + \frac{W}{4}$. But x_2 may not always be as large as x_1 for all values of x_1. This may be seen by taking $x_2 = x_1$ and $x_3 = x_4 = x_1 - W$ in (2). This leads to $x_1 = \bar{x} + \frac{W}{2}$. Thus, for $\bar{x} + \frac{W}{4} \le x_1 \le \bar{x} + \frac{W}{2}$, we see that x_1 is the upper limit on x_2. To determine the lower limit on x_2 for this region of variation of x_1 , we select x_2 as near $x_4 = x_1 - W$ as is possible without causing x_3 to exceed x_1 . But $x_3 = 4\bar{x} - 2x_1 - x_2 + W$. At most, then $4\bar{x} - 2x_1 - x_2 + W = x_1$ or $x_2 = 4\bar{x} - 3\bar{x}_1 + W$. Thus we have established the limits of integration used in the first part of the sum of which $F_1(\bar{x},W)$ consists. A similar argument shows if $\bar{x} + \frac{W}{2} \le x_1 \le \bar{x} + \frac{3W}{4}$, that

$$x_1 - W \le x_2 \le 4\bar{x} - 3x_1 + 2W.$$

If $f(x)$ is a probability function of the second kind, we observe in samples of three independent items x_1, x_2, x_3, for $\bar{x}$ assigned, that $0 \leq W \leq 3\bar{x}$. If $0 \leq W \leq 3\bar{x}/2$, we have

$$\bar{x} + \frac{W}{3} \leq x_1 \leq \bar{x} + \frac{2W}{3},$$
$$x_2 = 3\bar{x} - 2x_1 + W,$$
$$x_3 = x_1 - W,$$

and if $\frac{3\bar{x}}{2} \leq W \leq 3\bar{x}$, we have

$$W \leq x_1 \leq \bar{x} + \frac{2W}{3},$$
$$x_2 = 3\bar{x} - 2x_1 + W,$$
$$x_3 = x_1 - W.$$

Accordingly,

$$F_1(\bar{x}, W) = 18\int_{\bar{x}+\frac{W}{3}}^{\bar{x}+\frac{2W}{3}} f(x_1)f(x_1 - W)f(3\bar{x} - 2x_1 + W)dx_1, \quad 0 \leq W \leq \frac{3\bar{x}}{2},$$
$$= 18\int_{W}^{\bar{x}+\frac{2W}{3}} f(x_1)f(x_1 - W)f(3\bar{x} - 2x_1 + W)dx_1, \quad \frac{3\bar{x}}{2} \leq W \leq 3\bar{x}.$$

In samples of four independent items x_1, x_2, x_3, x_4, drawn from a universe characterized by a law of probability of this kind, we find

$$F_1(\bar{x}, W) = 48\int_{\bar{x}+\frac{W}{4}}^{\bar{x}+\frac{W}{2}} \int_{4\bar{x}-3x_1+W}^{x_1} f(x_1)f(x_2)f(4\bar{x} - 2x_1 - x_2 + W)f(x_1 - W)dx_2\,dx_1$$
$$+48\int_{\bar{x}+\frac{W}{2}}^{\bar{x}+\frac{3W}{4}} \int_{x_1-W}^{4\bar{x}-3x_1+2W} f(x_1)f(x_2)f(4\bar{x} - 2x_1 - x_2 + W)f(x_1 - W)dx_2\,dx_1,$$
$$0 \leq W \leq \frac{4\bar{x}}{3},$$
$$= 48\int_{W}^{\bar{x}+\frac{W}{2}} \int_{4\bar{x}-3x_1+W}^{x_1} f(x_1)f(x_2)f(4\bar{x} - 2x_1 - x_2 + W)f(x_1 - W)dx_2\,dx_1$$
$$+48\int_{\bar{x}+\frac{W}{2}}^{\bar{x}+\frac{3W}{4}} \int_{x_1-W}^{4\bar{x}-3x_1+2W} f(x_1)f(x_2)f(4\bar{x} - 2x_1 - x_2 + W)f(x_1 - W)\,dx_2\,dx_1,$$
$$\frac{4\bar{x}}{3} \leq W \leq 2\bar{x},$$
$$= 48\int_{W}^{\bar{x}+\frac{3W}{4}} \int_{x_1-W}^{4\bar{x}-3x_1+2W} f(x_1)f(x_2)f(4\bar{x} - 2x_1 - x_2 + W)f(x_1 - W)dx_2\,dx_1.$$
$$2\bar{x} \leq W \leq 4\bar{x}.$$

Finally, consider $f(x)$ to be a probability function of the third kind. In samples of three independent items x_1, x_2, x_3, for $0 \le \bar{x} \le k/3$, we obtain $0 \le W \le 3\bar{x}$; for $k/3 \le \bar{x} \le 2k/3$, we obtain $0 \le W \le k$; for $2k/3 \le \bar{x} \le k$, we obtain $0 \le W \le 3(k-\bar{x})$. It is fairly easy to see that for $\bar{x}$ and W assigned as indicated, the following regions of selection of x_1 are valid:

for $0 \le \bar{x} \le k/2$ and $0 \le W \le 3\bar{x}/2$,
or for $k/2 \le \bar{x} \le k$ and $0 \le W \le 3(k-\bar{x})/2$, then
$\bar{x} + W/3 \le x_1 \le \bar{x} + 2W/3$;
for $0 \le \bar{x} \le k/3$ and $3\bar{x}/2 \le W \le 3\bar{x}$,
or for $k/3 \le \bar{x} \le k/2$ and $3\bar{x}/2 \le W \le 3(k-\bar{x})/2$, then
$W \le x_1 \le \bar{x} + 2W/3$;
for $2k/3 \le \bar{x} \le k$ and $3(k-\bar{x})/2 \le W \le 3(k-\bar{x})$
or for $k/2 \le \bar{x} \le 2k/3$ and $3(k-\bar{x})/2 \le W \le 3\bar{x}/2$, then
$\bar{x} + W/3 \le x_1 \le k$;
for $k/3 \le \bar{x} \le k/2$ and $3(k-\bar{x})/2 \le W \le k$,
or for $k/2 \le \bar{x} \le 2k/3$ and $3\bar{x}/2 \le W \le k$, then $W \le x_1 \le k$.

Thus,

$$F_1(\bar{x},W) = 18\int_{\bar{x}+\frac{W}{3}}^{\bar{x}+\frac{2W}{3}} f(x_1)f(x_1-W)f(3\bar{x}-2x_1+W)dx_1,$$

$$=18\int_{W}^{\bar{x}+\frac{2W}{3}} f(x_1)f(x_1-W)f(3\bar{x}-2x_1+W)dx_1,$$

$$=18\int_{\bar{x}+\frac{W}{3}}^{k} f(x_1)f(x_1-W)f(3\bar{x}-2x_1+W)dx_1,$$

$$=18\int_{W}^{k} f(x_1)f(x_1-W)f(3\bar{x}-2x_1+W)dx_1,$$

over those regions of the $\bar{x}W$-plane indicated above.

In case of samples of four independent items x_1, x_2, x_3, x_4, drawn from a universe characterized by a probability function of the third kind, for $0 \le \bar{x} \le k/4$, we obtain $0 \le W \le 4\bar{x}$;

for $k/4 \leq \bar{x} \leq 3k/4$, we obtain $0 \leq W \leq k$; for $3k/4 \leq \bar{x} \leq k$, we obtain $0 \leq W \leq 4(k-\bar{x})$. Let us denote as follows the regions of the $\bar{x}W$-plane bounded by the given lines:

$$(A)\begin{cases} \bar{x} = 0 \\ W = \frac{4\bar{x}}{3} \\ W = \frac{4(k-\bar{x})}{3} \end{cases} \qquad (E)\begin{cases} W = \frac{4\bar{x}}{3} \\ W = 2(k-\bar{x}) \\ W = 4(k-\bar{x}) \end{cases}$$

$$(B)\begin{cases} W = \frac{4(k-\bar{x})}{3} \\ W = 2\bar{x} \\ W = \frac{4\bar{x}}{3} \end{cases} \qquad (F)\begin{cases} W = \frac{4\bar{x}}{3} \\ W = 2\bar{x} \\ W = \frac{4(k-\bar{x})}{3} \\ W = 2(k-\bar{x}) \end{cases}$$

$$(C)\begin{cases} W = 2\bar{x} \\ W = 4\bar{x} \\ W = \frac{4(k-\bar{x})}{3} \end{cases} \qquad (G)\begin{cases} W = k \\ W = 2\bar{x} \\ W = \frac{4(k-\bar{x})}{3} \end{cases}$$

$$(D)\begin{cases} W = \frac{4\bar{x}}{3} \\ W = 2(k-\bar{x}) \\ W = \frac{4(k-\bar{x})}{3} \end{cases} \qquad (H)\begin{cases} W = k \\ W = \frac{4\bar{x}}{3} \\ W = 2(k-\bar{x}) \end{cases}$$

Further, let

$$\theta = f(x_1)f(x_2)f(x_1-W)f(4\bar{x}-2x_1-x_2+W),$$

and let

$$\int_a^b\int_c^d \theta\, dx_2\, dx_1 = \begin{pmatrix} b & d \\ a & c \end{pmatrix} \theta.$$

It is then not difficult to verify that

$$F_1(\bar{x},W) = 48\left[\begin{pmatrix} \bar{x}+\frac{W}{2} & x_1 \\ \bar{x}+\frac{W}{4} & 4\bar{x}-3x_1+W \end{pmatrix}\theta + \begin{pmatrix} \bar{x}+\frac{3W}{4} & 4\bar{x}-3x_1+2W \\ \bar{x}+\frac{W}{2} & x_1-W \end{pmatrix}\theta\right], \quad (A)$$

$$= 48\left[\begin{pmatrix} \bar{x}+\frac{W}{2} & x_1 \\ W & 4\bar{x}-3x_1+W \end{pmatrix}\theta + \begin{pmatrix} \bar{x}+\frac{3W}{4} & 4\bar{x}-3x_1+2W \\ \bar{x}+\frac{W}{2} & x_1-W \end{pmatrix}\theta\right], \quad (B)$$

$$= 48\left[\begin{pmatrix} \bar{x}+\frac{3W}{4} & 4\bar{x}-3x_1+W \\ W & x_1-W \end{pmatrix}\theta\right], \quad (C)$$

$$= 48\left[\begin{pmatrix} \bar{x}+\frac{W}{2} & x_1 \\ \bar{x}+\frac{W}{4} & 4\bar{x}-3x_1+W \end{pmatrix}\theta + \begin{pmatrix} k & 4\bar{x}-3x_1+2W \\ \bar{x}+\frac{W}{2} & x_1-W \end{pmatrix}\right], \quad (D)$$

$$=48\left[\begin{pmatrix} k & x_1 \\ \bar{x}+\frac{W}{4} & 4\bar{x}-3x_1,+W \end{pmatrix}\right]\theta \quad , \tag{E}$$

$$=48\left[\begin{pmatrix} \bar{x}+\frac{W}{2} & x_1 \\ W & 4\bar{x}-3x_1,+W \end{pmatrix}\right]\theta \quad , \tag{F}$$

$$=48\left[\begin{pmatrix} k & 4\bar{x}-3x_1,+2W \\ W & x_1,-W \end{pmatrix}\right]\theta \; , \tag{G}$$

$$=48\left[\begin{pmatrix} k & x_1 \\ W & 4\bar{x}-3x_1,+W \end{pmatrix}\right]\theta \; . \tag{H}$$

As illustrations of these theorems, let us find the correlation between the range and the mean for universes of specified types.

Example 1. Let $f(x)=e,^{-x} \quad 0 \le x < \infty.$
For samples of three items, we have

$$F_1(\bar{x},W) = 6We^{-3\bar{x}}, \quad 0 \le W \le \frac{3\bar{x}}{2},$$

$$= 18\left(\bar{x}-\frac{W}{3}\right)e^{-3\bar{x}}, \quad \frac{3\bar{x}}{2} \le W \le 3\bar{x}.$$

The distributions of the marginal totals of W and $\bar{x}$ are obtained by integrating $F_1(\bar{x},W)$ with regard to $\bar{x}$ and W respectively. We readily find

$$\phi(\bar{x}) = \frac{27\bar{x}^2}{2}e^{-3\bar{x}}, \qquad 0 \le \bar{x} < \infty,$$

and

$$\psi(W) = 2e^{-2W}(e^{W}-1), \quad 0 \le W < \infty,$$

as previously given by the writer.[2] For $\bar{x}$ assigned, the mean of the array of W is $\overline{W}_{\bar{x}} = \frac{3\bar{x}}{2}$. Thus the regression of W on $\bar{x}$ is linear and $r = \frac{\sqrt{15}}{5}$.

[2] American Journal of Mathematics, Vol. 54 (1932), pp. 359, 366.

Example 2. Let $f(x) = 1/k, \quad 0 \leq x \leq k.$
For samples of three items, we have

$$F_1(\bar{x}, W) = \frac{6W}{k^3},$$

$$= \frac{18}{k^3}\left(\bar{x} - \frac{W}{3}\right),$$

$$= \frac{18}{k^3}\left(k - \bar{x} - \frac{W}{3}\right),$$

$$= \frac{18}{k^3}(k - W)$$

over those regions of the $\bar{x}W$-plane indicated above. The marginal totals[3] are distributed in accord with

$$\varphi(\bar{x}) = \frac{27\bar{x}^2}{2k^3}, \qquad 0 \leq \bar{x} \leq \frac{k}{3},$$

$$= \frac{9}{2k^3}\left[-6\bar{x}^2 + 6k\bar{x} - k^2\right], \quad \frac{k}{3} \leq \bar{x} \leq \frac{2k}{3},$$

$$= \frac{27}{2k^3}(k - \bar{x})^2, \quad \frac{2k}{3} \leq \bar{x} \leq k,$$

and $$\psi(W) = \frac{6W}{k^3}(k - W), \quad 0 \leq W \leq k.$$

We readily find

$$\overline{W}_{\bar{x}} = \frac{3\bar{x}}{2}, \quad 0 \leq \bar{x} \leq \frac{k}{3},$$

[3] Cf. H. L. Rietz, On a Certain Law of Probability of Laplace, Proc. Int. Math. Congress, Toronto (1924), pp. 795-799.

J. O. Irwin, On the Frequency Distributions of Means, etc., Biometrika, Vol. 19 (1927), pp. 225-239.

P. Hall, The Distribution of Means for Samples of Size N, Biometrika, Vol. 19 (1927), pp. 240-245.

J. Neyman and E. S. Pearson, On the Use and Distribution of Certain Test Criteria, Biometrika, Vol. 20 (1928), p. 210.

$$= \frac{5k^3 - 27k^2\bar{x} + 27k\bar{x}^2}{6k^2 - 36k\bar{x} + 36\bar{x}^2}, \quad \frac{k}{3} \le \bar{x} \le \frac{2k}{3},$$

$$= \frac{3}{2}(k-\bar{x}), \qquad \frac{2k}{3} \le \bar{x} \le k.$$

Thus the regression curve of W on $\bar{x}$ is continuous, but the regression is non-linear for $\frac{k}{3} \le \bar{x} \le \frac{2k}{3}$.

3. The correlation between the arithmetic mean $\bar{x}$ and the median ξ.

Theorem II. *Let $f(x)$ be the probability function of the variable x. Let $F_2(\bar{x}, \xi)$ be that of the arithmetic mean $\bar{x}$ and the median ξ in samples of three independent values of x. If $f(x)$ is a probability function of the first kind, then*

$$F_2(\bar{x},\xi) = 18 f(\xi)\int_{3\bar{x}-2\xi}^{\infty} f(x_1)\, f(3\bar{x}-\xi-x_1)\,dx_1, \quad \xi \le \bar{x},$$
$$= 18 f(\xi)\int_{\xi}^{\infty} f(x_1)\, f(3\bar{x}-\xi-x_1)\,dx_1, \quad \bar{x} \le \xi.$$

Proof. Let x_1, x_2, x_3, be the three observed values of x. Write

$$x_1 + x_2 + x_3 = 3\bar{x},$$
$$x_2 = \xi$$
$$x_3 \le x_2 \le x_1$$

For $\bar{x}$ and ξ assigned, $\xi \le \bar{x}$, we must have

$$3\bar{x} - 2\xi \le x_1 < \infty$$
$$x_2 = \xi$$
$$x_3 = 3\bar{x} - \xi - x_1,$$

and for $\bar{x} \le \xi$,

$$\xi \le x_1 < \infty$$
$$x_2 = \xi$$
$$x_3 = 3\bar{x} - \xi - x_1.$$

If we consider all possible arrangements of x_1 , x_2 , x_3 , we have

$$F_2(\bar{x},\xi)\,d\bar{x}\,d\xi = 6f(\xi)\,d\xi \int_{3\bar{x}-2\xi}^{\infty} f(x_1)\,f(x_3)\,dx_1\,dx_3, \qquad \xi \le \bar{x},$$

$$= 6f(\xi)\,d\xi \int_{\xi}^{\infty} f(x_1)\,f(x_3)\,dx_1\,dx_3, \qquad \bar{x} \le \xi.$$

The change of variable $x_3 = 3\bar{x} - \xi - x_1$ establishes the theorem.

In case of samples of five independent items x_1 , x_2 , x_3 , x_4 , x_5 , the probability function $F_2(\bar{x},\xi)$ is given by

$$F_2(\bar{x},\xi) = 150f(\xi)\int_{\xi}^{5\bar{x}-4\xi}\int_{5\bar{x}-3\xi-x_1}^{\infty}\int_{5\bar{x}-2\xi-x_1-x_2}^{\xi} f(x_1)f(x_2)f(x_4)f(5\bar{x}-\xi-x_1-x_2-x_4)\,dx_4\,dx_2\,dx_1$$

$$+150f(\xi)\int_{5\bar{x}-4\xi}^{\infty}\int_{\xi}^{\infty}\int_{5\bar{x}-2\xi-x_1-x_2}^{\xi} f(x_1)f(x_2)f(x_4)f(5\bar{x}-\xi-x_1-x_2-x_4)\,dx_4\,dx_2\,dx_1, \quad \xi \le \bar{x},$$

$$=150f(\xi)\int_{\xi}^{\infty}\int_{\xi}^{\infty}\int_{5\bar{x}-2\xi-x_1-x_2}^{\xi} f(x_1)f(x_2)f(x_4)f(5\bar{x}-\xi-x_1-x_2-x_4)\,dx_4\,dx_2\,dx_1, \quad \bar{x} \le \xi.$$

This follows immediately from the fact that for $\bar{x}$ and ξ assigned, $\xi \le \bar{x}$, we may have either

$$\begin{aligned} \xi \le x_1 &\le 5\bar{x} - 4\xi, \\ 5\bar{x} - 3\xi - x_1 \le x_2 &< \infty, \\ 5\bar{x} - 2\xi - x_1 - x_2 \le x_4 &\le \xi, \\ x_3 &= \xi, \\ x_5 &= 5\bar{x} - \xi - x_1 - x_2 - x_4, \end{aligned}$$

or

$$\begin{aligned} 5\bar{x} - 4\xi \le x_1 &< \infty, \\ \xi \le x_2 &< \infty, \\ 5\bar{x} - 2\xi - x_1 - x_2 \le x_4 &\le \xi \\ x_3 &= \xi \\ x_5 &= 5\bar{x} - \xi - x_1 - x_2 - x_4, \end{aligned}$$

and for $\bar{x} \le \xi$, we must have

$$\begin{aligned} \xi \le x_1 &< \infty \\ \xi \le x_2 &< \infty, \\ 5\bar{x}-2\xi-x_1-x_2 \le x_4 &\le \xi, \\ x_3 &= \xi \\ x_5 &= 5\bar{x}-\xi-x_1-x_2-x_4 . \end{aligned}$$

If $f(x)$ is a probability function of the second kind, it is clear that $0 \le \xi \le \frac{3\bar{x}}{2}$ in samples of three items. Then

$$F_2(\bar{x},\xi) = 18f(\xi)\int_{3\bar{x}-2\xi}^{3\bar{x}-\xi} f(x_1)f(3\bar{x}-\xi-x_1)dx_1 , \qquad 0 \le \xi \le \bar{x},$$

$$= 18f(\xi)\int_{\xi}^{3\bar{x}-\xi} f(x_1)f(3\bar{x}-\xi-x_1)dx_1 , \qquad \bar{x} \le \xi \le \frac{3\bar{x}}{2}.$$

In case of samples of five independent items drawn at random from a universe characterized by a probability function of the second kind, $F_2(\bar{x},\xi)$ can best be expressed in a form employing the notation used previously. Thus we write

$$\Phi = f(x_1)f(x_2)f(x_4)f(5\bar{x}-\xi-x_1-x_2-x_4),$$

$$u_{ij} = 5\bar{x}-i\xi-x_1-x_2-\cdots\cdots-x_j,$$

and

$$\int_a^b\int_c^d\int_e^f \Phi\, dx_4\, dx_2\, dx_1 = \begin{pmatrix} b & d & f \\ a & c & e \end{pmatrix}\Phi.$$

Then

$$F_2(\bar{x},\xi) = 150f(\xi)\left[\begin{pmatrix} u_{40} & u_{21} & \xi \\ \xi & u_{31} & u_{22} \end{pmatrix}\Phi + \begin{pmatrix} u_{40} & u_{11} & u_{12} \\ \xi & u_{21} & 0 \end{pmatrix}\Phi\right.$$

$$+ \begin{pmatrix} u_{30} & u_{21} & \xi \\ u_{40} & \xi & u_{22} \end{pmatrix}\Phi + \begin{pmatrix} u_{30} & u_{11} & u_{12} \\ u_{40} & u_{21} & 0 \end{pmatrix}\cdot\Phi$$

$$\left. + \begin{pmatrix} u_{20} & u_{11} & u_{12} \\ u_{30} & \xi & 0 \end{pmatrix}\Phi\right], \qquad 0 \le \xi \le \bar{x},$$

$$=150f(\xi)\left[\begin{pmatrix} u_{30} & u_{21} & \xi \\ \xi & \xi & u_{22} \end{pmatrix}\Phi + \begin{pmatrix} u_{30} & u_{11} & u_{12} \\ \xi & u_{21} & 0 \end{pmatrix}\Phi\right.$$

$$\left. + \begin{pmatrix} u_{20} & u_{11} & u_{12} \\ u_{30} & \xi & 0 \end{pmatrix}\Phi\right], \quad \bar{x} \le \xi \le \frac{5\bar{x}}{4},$$

$$=150 f(\xi)\left[\begin{pmatrix} u_{20} & u_{11} & u_{12} \\ \xi & \xi & 0 \end{pmatrix}\Phi\right], \quad \frac{5\bar{x}}{4} \le \xi \le \frac{5\bar{x}}{3}.$$

Finally, consider $f(x)$ to be a probability function of the third kind. In samples of three independent items, for $0 \le \bar{x} \le k/3$, we obtain $0 \le \xi \le 3\bar{x}/2$; for $k/3 \le \bar{x} \le 2k/3$, we obtain $(3\bar{x}-k)/2 \le \xi \le 3\bar{x}/2$; for $2k/3 \le \bar{x} \le k$, we obtain $(3\bar{x}-k)/2 \le \xi \le k$. It is not difficult to verify for $\bar{x}$ and ξ assigned as indicated, the following regions of selection of x_1 are valid:

for $0 \le \bar{x} \le k/3$ and $0 \le \xi \le \bar{x}$,
or for $k/3 \le \bar{x} \le k/2$ and $3\bar{x}-k \le \xi \le \bar{x}$, then
$3\bar{x}-2\xi \le x_1 \le 3\bar{x}-\xi$;
for $k/3 \le \bar{x} \le k/2$ and $(3\bar{x}-k)/2 \le \xi \le 3\bar{x}-k$,
or for $k/2 \le \bar{x} \le k$ and $(3\bar{x}-k)/2 \le \xi \le \bar{x}$, then
$3\bar{x}-2\xi \le x_1 \le k$;
for $0 \le \bar{x} \le k/2$ and $\bar{x} \le \xi \le 3\bar{x}/2$,
or for $k/2 \le \bar{x} \le 2k/3$ and $3\bar{x}-k \le \xi \le 3\bar{x}/2$, then
$\xi \le x_1 \le 3\bar{x}-\xi$;
for $k/2 \le \bar{x} \le 2k/3$ and $\bar{x} \le \xi \le 3\bar{x}-k$,
or for $2k/3 \le \bar{x} \le k$ and $\bar{x} \le \xi \le k$, then $\xi \le x_1 \le k$.

Thus

$$F_2(\bar{x},\xi) = 18f(\xi)\int_{3\bar{x}-2\xi}^{3\bar{x}-\xi} f(x_1)f(3\bar{x}-\xi-x_1)\,dx_1,$$

$$= 18f(\xi)\int_{3\bar{x}-2\xi}^{k} f(x_1)f(3\bar{x}-\xi-x_1)\,dx_1,$$

$$= 18 f(\xi) \int_{\xi}^{3\bar{x}-\xi} f(x_1)\, f(3\bar{x}-\xi-x_1)\,dx_1,$$

$$= 18 f(\xi) \int_{\xi}^{k} f(x_1)\, f(3\bar{x}-\xi-x_1)\,dx_1,$$

over those regions of the $\bar{x}\xi$-plane as indicated above.

With samples of five items, the correlation surface is defined in so many parts that we shall not take the space necessary to consider it.

As illustrations of these theorems, we shall find the correlation between the median and the mean for universes of specified types.

Example 1. Let $f(x) = e^{-x}, \quad 0 \le x < \infty.$
For samples of three items, we have

$$F_2(\bar{x},\xi) = 18\xi e^{-3\bar{x}}, \quad 0 \le \xi \le \bar{x},$$
$$= 18(3\bar{x}-2\xi)e^{-3\bar{x}}, \quad \bar{x} \le \xi \le \frac{3\bar{x}}{2}$$

The distribution function of the marginal totals of ξ is given by[4]

$$\varphi(\xi) = 6e^{-2\xi}(1-e^{-\xi}), \quad 0 \le \xi < \infty.$$

For $\bar{x}$ assigned, the mean of the array of ξ is

$$\bar{\xi}_{\bar{x}} = \frac{5\bar{x}}{6}.$$

Thus the regression of ξ on $\bar{x}$ is linear and $r = \frac{5\sqrt{267}}{89}.$

Example 2. Let $f(x) = \frac{1}{k}, \quad 0 \le x \le k.$
For samples of three items, we have

$$F_2(\bar{x},\xi) = \frac{18\xi}{k^3}$$

$$= \frac{18}{k^3}(k-3\bar{x}+2\xi),$$

[4] Cf. American Journal of Mathematics, Vol. 54 (1932), p. 364.

$$= \frac{18}{k^3}(3\bar{x}-2\xi),$$

$$= \frac{18}{k^3}(3k-\xi),$$

over those regions of the $\bar{x}\,\xi$-plane indicated above. The distribution function of the marginal totals of ξ is given by[5]

$$\phi(\xi) = \frac{6\xi}{k^3}(k-\xi), \qquad 0 \leq \xi \leq k.$$

We find

$$\bar{\xi}_{\bar{x}} = \frac{5\bar{x}}{6}, \qquad 0 \leq \bar{x} \leq \frac{k}{3},$$

$$= \frac{5\bar{x}^3-(3\bar{x}-k)^3}{6\bar{x}^2-2(3\bar{x}-k)^2}, \qquad \frac{k}{3} \leq \bar{x} \leq \frac{2k}{3},$$

$$= \frac{(5\bar{x}+k)}{6}, \qquad \frac{2k}{3} \leq \bar{x} \leq k.$$

Thus the regression curve of ξ on $\bar{x}$ is continuous but the regression is non-linear for $\frac{k}{3} \leq \bar{x} \leq \frac{2k}{3}$..

4. The correlation between the median ξ and the range W.

Theorem III. *Let $f(x)$ be the probability function of the variable x. Let $F_3(\xi,W)$ be that of the median ξ and the range W in samples of $2m+1$ independent values of x. If $f(x)$ is a probability function of the first kind, then*

$$F_3(\xi,W) = \frac{(2m+1)!}{[(m-1)!]^2} f(\xi) \int_{\xi}^{\xi+W} f(x_1)f(x_1-W)\left[\int_{\xi}^{x_1} f(t)dt\right]^{m-1}\left[\int_{x_1-W}^{\xi} f(t)dt\right]^{m-1} dx_1.$$

Proof. We have

$$x_1 - x_{2m+1} = W,$$

$$x_{m+1} = \xi,$$

$$\xi \leq x_2, \cdots\cdots, x_m \leq x_1,$$

$$x_1 - W \leq x_{m+1}, \cdots\cdots, x_{2m} \leq \xi.$$

[5] Cf. P. R. Rider, On the Distribution of the Ratio of Mean to Standard Deviation, etc., Biometrika, Vol. 21 (1929), pp. 136-137.

Hence the theorem.

If $f(x)$ is a probability function of the second kind, then

$$F_3(\xi,W)=\frac{(2m+1)!}{[(m-1)!]^2}f(\xi)\int_{\xi}^{\xi+W}f(x_1)f(x_1-W)\left[\int_{\xi}^{x_1}f(t)dt\right]^{m-1}\left[\int_{x_1-W}^{\xi}f(t)dt\right]^{m-1}dx_1,\ W\le\xi,$$

$$=\frac{(2m+1)!}{[(m-1)!]^2}f(\xi)\int_{W}^{\xi+W}f(x_1)f(x_1-W)\left[\int_{\xi}^{x_1}f(t)dt\right]^{m-1}\left[\int_{x_1-W}^{\xi}f(t)dt\right]^{m-1}dx_1,\ \xi\le W.$$

Finally, consider $f(x)$ to be a probability function of the third kind. We observe for $0\le\xi\le k$, that $0\le W\le k$. For assigned values of ξ and W, the following regions of selection of x_1 are obvious:

for $0\le\xi\le k/2$, and $0\le W\le\xi$,
or for $k/2\le\xi\le k$ and $0\le W\le k-\xi$, then $\xi\le x_1\le\xi+W$;
for $0\le\xi\le k/2$ and $\xi\le W\le k-\xi$, then $W\le x_1\le\xi+W$;
for $0\le\xi\le k/2$, and $k-\xi\le W\le k$,
or for $k/2\le\xi\le k$ and $\xi\le W\le k$, then $W\le x_1\le k$;
for $k/2\le\xi\le k$ and $k-\xi\le W\le\xi$, then $\xi\le x_1\le k$.

If we write

$$\psi=f(x_1)f(x_1-W)\left[\int_{\xi}^{x_1}f(t)dt\right]^{m-1}\left[\int_{x_1-W}^{\xi}f(t)dt\right]^{m-1},$$

we have

$$F_3(\xi,W)=\frac{(2m+1)!}{[(m-1)!]^2}f(\xi)\int_{\xi}^{\xi+W}\psi\,dx_1,$$

$$=\frac{(2m+1)!}{[(m-1)!]^2}f(\xi)\int_{W}^{\xi+W}\psi\,dx_1,$$

$$=\frac{(2m+1)!}{[(m-1)!]^2}f(\xi)\int_{W}^{k}\psi\,dx_1,$$

$$=\frac{(2m+1)!}{[(m-1)!]^2}f(\xi)\int_{\xi}^{k}\psi\,dx_1,$$

over those regions of the ξW-plane previously indicated.

We shall consider two simple examples.

Example 1. Let $f(x) = e^{-x}, \quad 0 \leq x < \infty.$
With samples of three items,

$$F_3(\xi, W) = 3e^{-3\xi}(e^{W} - e^{-W}), \quad W \leq \xi,$$
$$= 3e^{-W-\xi}(1-e^{-2\xi}), \quad \xi \leq W.$$

The regression is readily shown to be non-linear.

Example 2. Let $f(x) = \frac{1}{k}, \quad 0 \leq x \leq k.$
With samples of three items,

$$F_3(\xi, W) = \frac{6W}{k^3},$$
$$= \frac{6\xi}{k^3},$$
$$= \frac{6}{k^3}(k-W)$$
$$= \frac{6}{k^3}(k-\xi),$$

over those regions of the ξW-plane which have been previously given. The mean of the array of W corresponding to an assigned ξ is $\overline{W}_{\xi} = \frac{k}{2}$. Accordingly, there is no correlation between the median and the range in samples of three items drawn from this universe.

It is easy to employ the type of argument used in establishing Theorem III to obtain the probability function of the median and lower quartile. Thus, if $f(x)$ is a probability function of the second kind and $F_4(\xi, \eta)$ is the probability function of the median ξ and the lower quartile η in samples of $4m+1$ items, then

$$F_4(\xi, \eta) = \frac{(4m+1)!}{(2m)!\,m!\,(m-1)!} f(\xi) f(\eta) \left[\int_{\xi}^{\infty} f(t)dt\right]^{2m} \left[\int_{0}^{\eta} f(t)dt\right]^{m}$$
$$\cdot \left[\int_{\eta}^{\xi} f(t)dt\right]^{m-1}, \qquad n \leq \xi.$$

Allen T. Craig

ON THE DEGREE OF APPROXIMATION OF CERTAIN QUADRATURE FORMULAS

By

A. L. O'Toole
National Research Fellow.

If $f(x)$ be a continuous function of period 2π, and if the interval under consideration, say the interval from 0 to 2π, be divided into m equal parts by the $m+1$ points $x_i = 2i\pi/m$, $i=0, 1, 2, \ldots\ldots, m$, then the trigonometric sum of the nth order coinciding in value with $f(x)$ at the $m+1$ points x_i, or the trigonometric sum of the nth order lacking the term in sin nx, is, according as $m = 2n+1$ or $m=2n$,

$$\begin{aligned}\Phi_n(x) = \tfrac{1}{2}a_0 + a_1 \cos x + a_2 \cos 2x + \cdots\cdots + a_n \cos nx \\ + b_1 \sin x + b_2 \sin 2x + \cdots\cdots + b_n \sin nx\end{aligned}$$

or

$$\begin{aligned}u_n(x) = \tfrac{1}{2}a_0 + a_1 \cos x + a_2 \cos 2x + \cdots\cdots + \tfrac{1}{2}a_n \cos nx \\ + b_1 \sin x + b_2 \sin 2x + \cdots + b_{n-1} \sin(n-1)x,\end{aligned}$$

where

$$a_k = \frac{h}{\pi}\sum_{i=1}^{m} f(x_i)\cos kx_i . \qquad h = \frac{2\pi}{m} .$$

$$b_k = \frac{h}{\pi}\sum_{i=1}^{m} f(x_i)\sin kx_i .$$

If the Fourier coefficients of $f(x)$ be denoted by

$$\alpha_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos kx\, dx,$$

$$\beta_k = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin kx\, dx.$$

then it has been shown[1] that the interpolating coefficients a_k and b_k are approximations to the Fourier coefficients α_k and β_k in the sense of the rectangle quadrature formula, in the sense of the trapezoid quadrature formula, in the sense of the average of the results of two applications of Simpson's formula, and in the sense of higher quadrature formulas. In other words, the simple rectangle formulas a_k and b_k are as good approximations to the areas α_k and β_k as the estimates given by the trapezoid rule, the average of two applications of Simpson's rule, or higher quadrature formulas.

It is the purpose of this note to discuss certain quadrature formulas and to observe some other conditions under which the rectangle formula will give as good an approximation as the more complicated formulas.

The most elementary and best known of the formulas are the rectangle formula, the trapezoid formula, and Simpson's formula. Many of the more complex rules are the results of attempts by different investigators[2] to improve by various devices the approximations given by these three simple rules.

Suppose the area under consideration is bounded by the curve $y = f(x)$, the x-axis and the ordinates at $x = a$ and $x = b$. If the interval from a to b be devided into n equal[3] parts, say of length h, by the $n+1$ points $x_0 = a, x_1, x_2, \cdots, x_{n-1}, x_n = b$, and if rectangles, each of width h and height y_i, $i = 0, 1, 2, \cdots, n-1$, be constructed, then the area as approximated by these n rectangles is

$$(1) \qquad A = h \sum_{v=0}^{n-1} y_v .$$

[1] D. Jackson, Some Notes on Trigonometric Interpolation, Amer. Math. Monthly, vol. xxxiii, no. 8, October 1927.

[2] See Runge and Willers, Encyklopädie Der Mathematischen Wissenschaften, Bd. II:3 (1915), pp. 45-176.

[3] Discussion from point of view of least squares, Otto Biermann, Monatshefte Fur Mathematik Und Physik, 14 (1903), pp. 226-242.

For unequal intervals see Jas. W. Glover, International Mathematical Congress, Toronto, 1924.

To find an expression for the error we assume the first derivative exists, so that for the first rectangle

$$f(x) = f(a) + (x-a)f'(u),$$

$$\int_a^{a+h} f(x)\,dx = hf(a) + \frac{h^2}{2} f'(z), \quad a < z < a+h.$$

Hence the error for the n rectangles is

$$E = \frac{h^2}{2} \sum_{v=1}^{n} f'(z_v) = \frac{(b-a)^2}{2n} f'(z), \quad a < z < b, \tag{1e}$$

i.e., an error of the order of $\frac{1}{n}$.

Let $n = mk$, $k = 1, 2, 3, \cdots$. If we approximate the area in the first k subintervals by a parabola of degree k coinciding in value with $f(x)$ at the first k values of x, then integrating Lagrange's interpolation formula an expression for the error is obtained. If k is odd then

$$E_1 = C_1 \left(\frac{H}{2}\right)^{k+2} \frac{f^{(k+1)}(z)}{(n+1)!}, \quad H = kh,$$

where

$$C_1 = \frac{1}{(k+2)k^{k-1}} \int_{-1}^{1} (t^2-1)(k^2t^2-1^2)(k^2t^2-3^2)\cdots\cdots(k^2t^2-(k-2)^2)\,dt.$$

If k is even, then making use of Rolle's Theorem,

$$E_2 = C_2 \left(\frac{H}{2}\right)^{k+3} \frac{f^{(k+2)}(z)}{k+2!}$$

where

$$C_2 = \frac{1}{k^{k-2}} \int_{-1}^{1} (t^2-1)(t^2)(k^2t^2-2^2)\cdots(k^2t^2-(k-2)^2)\,dt.$$

The error over the whole interval will be obtained by summing the m errors corresponding to each k subintervals.

If n trapezoids are formed by joining the ends of successive ordinates then the area as approximated by the sum of the areas of these trapezoids is

(2) $$A=\frac{h}{2}\sum_{v=0}^{n-1}(y_v+y_{v+1})$$

and the error is

(2e) $$E=-\frac{(b-a)^3}{12n^2}f''(z),$$

i.e., an error of the order of $\frac{1}{n^2}$.

Simpson's formula may be obtained by passing second degree parabolas through the ends of three successive ordinates, that is $k=2$, and gives

(3) $$A=\frac{h}{3}\left[2\sum_{v=0}^{m}y_{2v}+4\sum_{v=1}^{m}y_{2v-1}-(y_0+y_{2m})\right],\ n=2m.$$

The error is

(3e) $$E=-\frac{(b-a)^5}{180n^4}f^{iv}(z),$$

i.e., an error of the order of $\frac{1}{n^4}$.

To illustrate the fact that sometimes the rectangle formula (1) gives a better approximation than the Simpson formula (3) these formulas will be applied to the problem of finding the area under the so-called normal curve of error. From a table[4] giving five places of decimals it is seen that the ordinates to the right of $x=4.76$ and to the left of $x=-4.76$ are everywhere zero if the equation be written in the form $y=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. Divide the interval from $x=-4.80$ to $x=4.80$ into eight partial intervals each of length 1.20. Formula (1) gives $A=.99998$ while

[4] Jas. W. Glover, Tables of Applied Mathematics in Finance, Insurance and Statistics.

(3) gives $A = .97834$, the same ordinates being used in each case.

There are three objections to the nature of Simpson's formula. They are the lack of smoothness at the points of intersection of the parabolas, the unequal weights attached to the odd and even numbered ordinates, and the requirement that the number of ordinates be odd.

Catalan[6] notices the lack of smoothness at the intersections of the parabolas used in setting up Simpson's rule and improves on it by passing parabolas through three successive ordinates and then retaining only the first half of each parabola except in the case of the last three ordinates where it is necessary to retain the whole parabola. To counterbalance the asymmetry introduced by these last three ordinates he repeats the process beginning with the last ordinat and then takes the arithmetic mean of the two results as his formula.
This gives

$$(4) \quad A = h\left[\sum_{v=0}^{n} y_v - \frac{5}{8}(y_0+y_n) + \frac{1}{6}(y_1+y_{n-1}) - \frac{1}{24}(y_2+y_{n-2})\right].$$

And, of course, the error is still of the order of $\frac{1}{n^4}$. This formula has the additional advantage that it holds no matter whether n is even or odd.

Similarly Crotti[6] showed that the different weights attached to the odd and even numbered ordinates in Simpson's formula ic a disadvantage. And Parmentier[7] by subtracting Simpson's formula from twice Catalan's obtained a formula in which the weights are the reverse of those in Simpson's. Mansion[8] gave an alternative derivation of Catalan's formula, his derivation requiring, however, an even number of ordinates.

[5] E. Catalan, Nouvelles Annales, 1re series (1851), pp. 412-415.

[6] Crotti, Il Politechnio 33 (1885), pp. 193-207.

[7] Parmentier, Association française pour l'avancement des sciences, Session Grenoble, 1882.

[8] Mansion, Supplement zu Mathesis 1 (1881).

Catalan's formula may be thought of as the 'rectangle formula plus three correctional terms involving the first three and the last three ordinates. In the case of an even number of ordinates a formula[9] involving only two such correctional terms and giving an approximation of the order of the error in the single trapezoid, i.e., of the order of $\frac{1}{n^3}$, the error in a single trapezoid of width $\frac{(b-a)}{n}$ being $\frac{-f''(z)(b-a)^3}{12n^3}$, can be obtained by applying Simpson's formula to the first $2m-1$ ordinates and approximating the remaining area by the trepezoid rule. Repeat the process from the opposite end and take the arithmetic mean of the two results as the quadrature formula. This gives

$$(5) \qquad a = h\left[\sum_{v=0}^{n} y_v - \frac{7}{12}(y_0+y_n) + \frac{1}{12}(y_1+y_{n-1})\right], \quad n=2m-1.$$

It is the only formula with just two correctional terms which will give even this order of approximation in general because any change in the coefficients of these end ordinates will introduce in general an error of the order of the error in the rectangle formula for a single subinterval, i.e., an error of the order of $\frac{1}{n^2}$.

Another important quadrature formula is called the three-eighths rule and is obtained by passing third order parabolas through four successive ordinates. It may be written

$$(6) \qquad A = \frac{3h}{8}\left[2\sum_{v=0}^{m} y_{3v} + 3\sum_{v=0}^{m-1} y_{3v+1} + 3\sum_{v=0}^{m-1} y_{3v+2} - (y_0+y_{3m})\right], \quad n=3m.$$

The error is

$$(6e) \qquad E = -\frac{(b-a)^5}{400n^4} f^{iv}(z),$$

i.e., an error of the same order as the error corresponding to Simpson's formula. The error terms derived from the Lagrange

[9] Durand, Engineering News, Jan. 1894. J. Lipka, Graphical and Mechanical Computation, Part II, p. 226.

formula shows the advantage of using parabolas of even degree.

Besides the fact that the order of the error is the same as that in the case of Simpson's formula, this three-eighths formula has disadvantages similar to those mentioned in the case of Simpson's formula. There is still a lack of smoothness at the intersections of the parabolas; the weights attached to the ordinates are as undesirable as before; and the number of partial intervals must be a multiple of three.

It is possible however to do away with these disadvantages by proceeding as follows. Pass a third order parabola through the first four ordinates y_0, y_1, y_2, y_3. Retain only the area in the first two partial intervals. Pass a third order parabola through the four ordinates y_1, y_2, y_3, y_4 and retain only the area in the central interval. Proceed in this way retaining each time only the area in the central interval until the last four ordinates are reached where it will again be necessary to retain the area in two strips, viz., the last two partial intervals. The sum of these areas gives the required quadrature formula. It is

$$(7)\ A=h\left[\sum_{v=0}^{n} y_v-\frac{2}{3}(y_0+y_n)+\frac{7}{24}(y_1+y_{n-1})-\frac{1}{6}(y_2+y_{n-2})+\frac{1}{24}(y_3+y_{n-3})\right].$$

This formula holds for any n greater than or equal to three. From the point of view of the order of the error this formula is, as one would expect, no better than Catalan's formula. As a matter of fact formula (7) can be obtained from formula (4) by subtracting from (4) $\frac{h}{24}(\Delta^3 y_0-\Delta^3 y_{n-3})$ a quantity which, in general, is of the order of $\frac{1}{n^4}$.

If $n=4m$ and fourth order parabolas are used in approximating the area in four successive partial intervals then the formula is

$$(8)\ A=\frac{4h}{45}\left[7\sum_{v=0}^{m} y_{4v}+16\sum_{v=0}^{m-1} y_{4v+1}+6\sum_{v=0}^{m-1} y_{4v+2}+16\sum_{v=0}^{m-1} y_{4v+3}-\frac{7}{2}(y_0+y_{4m})\right].$$

The error is

$$E = -\frac{2(b-a)^7}{945n^6} f^{vi}(z), \tag{8e}$$

i.e., an error of the order of $\frac{1}{n^6}$.

Several modifications may be made to improve this formula. For instance if $n=2m+1$ then apply the fourth degree parabola to the ordinates y_0 , y_1 , y_2 , y_3 , y_4 and retain only the area in the first three strips. Apply a fourth degree parabola to the ordinates y_2 , y_3 , y_4 , y_5 , y_6 and retain the area in the two central strips. And so on till in the final step it will be necessary to retain the area in the last three strips. Addition gives the formula

$$A=\frac{h}{720}\Big[896\sum_{v=0}^{m} y_{2v} + 544\sum_{v=1}^{m} y_{2v-1} - 653(y_0+y_{2m}) + 374(y_1+y_{2m-1})$$
$$-256(y_2+y_{2m-2}) + 106(y_3+y_{2m-3}) - 19(y_4+y_{2m-4})\Big] \tag{9}$$

A formula which holds for any n may be obtained by passing a fourth degree parabola through y_0 , y_1 , y_2 , y_3 , y_4 and retaining only the area between y_0 and y_2 . Pass a fourth degree parabola through y_1 , y_2 , y_3 , y_4 , y_5 and retain only the area between y_2 and y_3 . And so on, retaining only the area in one strip, until at the end it will be necessary to retain the area in the last three strips. Repeat the process beginning at the last ordinate and take the arithmetic mean. The result is

$$A = h\Big[\sum_{v=0}^{n} y_v - \frac{193}{288}(y_0+y_n) + \frac{77}{240}(y_1+y_{n-1}) - \frac{7}{30}(y_2+y_{n-2})$$
$$+\frac{73}{720}(y_3+y_{n-3}) - \frac{3}{160}(y_4+y_{n-4})\Big]. \tag{10}$$

This formula can be obtained in the case of an even number of ordinates by retaining three strips at the beginning, two from then on, reversing the process and taking the arithmetic mean.

Formulas (4), (5), (7) and (10) not only give, in general, at least as good approximations as Simpson's formula, the trapezoid formula, the three-eighths formula, and the fourth degree formula (8) respectively, but in addition have the important property that under certain conditions they show that the simple rectangle formula must give at least as good an approximation as the higher formulas. If $f(x)$ is a function such that the curve $y = f(x)$ actually, or at least for practical purposes, coincides with the x -axis to the left of $x=a$ and to the right of $x=b$, then in dividing the interval from a to b into h equal parts each of length h it will not affect the area required if two, one, three or four partial intervals of length h are marked off to the left of a and to the right of b, the number of such partial intervals corresponding to (4), (5), (7) and (10) respectively. Hence it is seen that under these conditions (4), (5), (7) and (10) reduce to the simple rectangle formula (1).

If the curve coincides with the x -axis at one end of the interval over which the area is required but does not at the other end then formulas (4), (5), (7) and (10) become respectively

$$(4a)\quad A=h\left(\sum_{v=0}^{n} y_v - \frac{5}{8}y_n + \frac{1}{6}y_{n-1} - \frac{1}{24}y_{n-2}\right),$$

$$(5a)\quad A=h\left(\sum_{v=0}^{2m-1} y_v - \frac{7}{12}y_{2m-1} + \frac{1}{12}y_{2m-2}\right),$$

$$(7a)\quad A=h\left(\sum_{v=0}^{n} y_v - \frac{2}{3}y_n + \frac{7}{24}y_{n-1} - \frac{1}{6}y_{n-2} + \frac{1}{24}y_{n-3}\right),$$

$$(10a)\quad A=h\left(\sum_{v=0}^{n} y_v - \frac{193}{288}y_n + \frac{77}{240}y_{n-1} - \frac{7}{30}y_{n-2} + \frac{73}{720}y_{n-3} - \frac{3}{160}y_{n-4}\right).$$

For example, consider again the normal curve of error and suppose that the area to the left of the ordinate at $x=o$ is required. Formulas (4a), (5a), (7a) and (10a) apply and for sixteen partial intervals give respectively $A=.49994$, $A=.49550$, $A=.50008$, and $A=.50002$, an extra partial interval to the left of $x=-4.80$ being used in the case of (5a) in order to have an odd number of intervals for that formula. Using thirty-two partial intervals the same formulas give $A=.49999$, $A=.49949$, $A=.50000$, and $A=.50000$ respectively.

If, as often happens, the values of ordinates outside the interval over which the area is required are known then even better quadrature formulas may be obtained. For example, suppose that in deriving formula (7) the ordinate y_{-1} at a distance of h to the left of y_0 and the ordinate y_{n+1} at a distance h to the right of y_n are known. Then it will not be necessary to retain the areas in double strips at the beginning and end of the interval, and the formula for the area over the interval from $x=a$ to $x=b$ is

$$(11)\quad A=h\left[\sum_{\nu=0}^{n} y_\nu - \frac{1}{24}(y_{-1}+y_{n+1}) - \frac{1}{2}(y_0+y_n) + \frac{1}{24}(y_1+y_{n-1})\right].$$

It should be noted that in case y_{-1} and y_{n+1} are known Catalan's formula reduces to (11). And, similarly, in the case of the derivation of formula (10) it will be necessary to retain the area in a single strip each time except in the case of the last application of the fourth degree parabola when it will be necessary to retain the area in the two central strips. The formula arrived at is

$$(12)\quad A=h\left[\sum_{\nu=0}^{n} y_\nu - \frac{3}{160}(y_{-1}+y_{n+1}) - \frac{83}{144}(y_0+y_n) + \frac{2}{15}(y_1+y_{n-1}) - \frac{11}{240}(y_2+y_{n-2}) + \frac{11}{1440}(y_3+y_{n-3})\right].$$

Formulas (11) and (12) reduce to the rectangle formula (1) under the same conditions as in the cases of (4), (5), (7) and (10). Likewise when the curve coincides with the x-axis to the left of $x=a$ (11) and (12) become

(11a) $A = h\left(\sum_{v=0}^{n} y_v - \frac{1}{24} y_{n+1} - \frac{1}{2} y_n + \frac{1}{24} y_{n-1}\right)$, and

(12a) $A = h\left(\sum_{v=0}^{n} y_v - \frac{3}{160} y_{n+1} - \frac{83}{144} y_n + \frac{2}{15} y_{n-1} - \frac{11}{240} y_{n-2} + \frac{11}{1440} y_{n-3}\right)$

If we apply formula (11a) to finding the area under the normal curve to the left of the ordinate at $x=0$ and take $n=4$, $h=1.20$, $a=-4.80$ then we find $A=.49999$. In other words, in this case (11a) gives as good a result with six ordinates as (4a) or (7a) give with thirty-three ordinates or (6a) with thirty-four ordinates.

Quadrature formulas involving parabolas of degree higher than four have been obtained but they are to be used with caution on account of the great freedom they allow the approximating curves. However, modifications similar to those in this paper could also be made for these higher formulas. And the effect of any number of ordinates outside the ends of the interval could be noted.

This note will be concluded with a remark on the effect of errors in the data giving the values of the ordinates. Suppose the quadrature formula is $A = h(a_0 y_0 + a_1 y_1 + a_2 y_2 + \cdots\cdots + a_n y_n)$ and suppose further that each y_i is subject to an error e_i, $i=0,1,2,3,\cdots,n$. If e is the greatest of the absolute values of the e_i then the error in A cannot be greater than $he(a_0+a_1+a_2+\cdots+a_n)$ if $a_0, a_1, a_2, \cdots\cdots, a_n$ are all positive, as will be true if parabolas of the fourth degree or lower are used. But $h(a_0+a_1+a_2+\cdots+a_n)=(b-a)$ if the area is to be four from $x=a$ to $x=b$. Hence the error in A due to errors in the data is not greater than $e(b-a)$. When parabolas of degree higher than four are used the coefficients in the quadrature formula are not always positive.

A. L. O'Toole